A Statistical-Physics Approach to Language Acquisition and Language Change ⋆

Marzio Cassandro

Dipartimento di Fisica, Università di Roma La Sapienza
Piazzale Aldo Moro, 5, OO185 Roma, Italy, cassandro@roma1.infn.it

Pierre Collet

CNRS and Centre de Physique Théorique, École Polytechnique
91128 Palaiseau Cedex, France, collet@cpht.polytechnique.fr

Antonio Galves [1]

Instituto de Matemática e Estatística, Universidade de São Paulo
BP 66281, 05315-970 São Paulo, SP, Brasil, galves@ime.usp.br

Charlotte Galves [2]

Instituto de Estudos da Linguagem, Universidade Estadual de Campinas
BP 6045, 13081-970, Campinas, SP, Brasil, galvesc@iel.unicamp.br

Abstract

The aim of this paper is to explain why Statistical Physics can help understanding two related linguistic questions. The first question is how to model first language acquisition by a child. The second question is how language change proceeds in time.

Our approach is based on a Gibbsian model for the interface between syntax and prosody. We also present a simulated annealing model of language acquisition, which extends the Triggering Learning Algorithm recently introduced in the linguistic litterature.

Key words: grammar identification, language change, Gibbs states, maximum likelihood, entropy, simulated annealing

1 Introduction

The aim of this paper is to explain why Statistical Physics can help understanding two related linguistic questions. The first question is how to model first language acquisition by a child. The second question is how language change proceeds in time.

We shall show that the Thermodynamical Formalism provides a suitable framework, in which the notions of prosodic pattern and syntax can be put together. They will both appear in the definition of the probability measure governing the choice of the sample of positive evidence offered to a child during the process of acquisition of his mother tongue. We propose to define this probability measure as a Gibbs state in which the prosody is decribed by the potential and the syntax is described by algebraic restricitons on the set of possible configurations. Roughly speaking, given a discursive context, the syntax says which sentences are available and prosody says what is the probability to choose a sentence among all the available ones.

With this model it is possible to formulate precisely the structure recognition procedure which is behind language acquisition. Our model accounts for the robustness of language acquisition even in the presence of a restricted sample of sentences provided as positive evidence. It also accounts for language change. Depending on the prosodic pattern, the identification procedure may lead the learning child to chose a grammar which differs from the parental one.

This paper is organized as follows. In section 2, the linguistic theoretical framework is briefly sketched. In section 3, we present a Gibbsian model for the interface between syntax and prosody. In section 4, we discuss the relationship between structure recognition and language change. As an application, we discuss a concrete case of language change, the one which leads from Classical to Modern European Portuguese. Finally, in section 5, a simulated annealing model of language acquisition is introduced, as an extension of the Triggering Learning Algorithm recently introduced in the linguistic litterature.

⋆ Expanded version of a talk presented by A.G. at StatPhys20. Work supported by FAPESP (Projeto Temático *Rhythmic patterns, parameter setting and language change*, grant 98/3382-0)

[1] Work partially supported CNPq (grant 301301/79) and FINEP (Núcleo de Excelência *"Critical phenomena in probability and stochastic processes"*

[2] Work partially supported by CNPq (grant 301086/85)

2 Linguistic framework

The view of language that has developed in modern linguistics is that there is a genetic inherited *linguistic capacity* which makes children able to learn a language. This linguistic capacity is characterized by a finite set of constraints, what Chomsky calls the *Universal Grammar*. Any particular solution of these constraints is called a *grammar* and defines in a precise way a natural language. Moreover, each grammar can be defined in terms of a finite (and not large) number of parameters each of which takes one of a small number of values.

As a consequence of this, first language acquisition is a process in which these parameters are set in a particular way. To do this the learning child assigns syntactic structures to the ordered strings of words he is exposed to as positive evidence from the parental language. The values of the parameters are identified when a structure is assigned to each one of the strings of words present in a sample rich enough.

To make guesses about the structures of the sentences of the sample the child takes advantage of the hints which are provided by the parental prosody (cf.[20,21]). Informally speaking, the prosody of a language is its characteristic *music*, which contains, among other things, its typical stress and intonational patterns.

It is important to emphasize that most psycholinguists agree that first language acquisiton is an unsupervised learning process, in the sense that only positive evidence is taken into account by the child (c.f. [19]).

Language change takes place when a generation of learning children assign structures which are different from the ones assigned by their parents to some of the sentences in the sample. It has been argued that some of these changes may have been induced by a change of the prosodic patterns of the language. Our model provides a clear picture of this effect.

The point of view we adopt here is based on the *Principles and Parameters* approach to grammar which has been developed by Chomsky and collaborators in the last three decades. For an up-to-date presentation of this approach we refer the reader to [9].

The present paper is part of a research project at the interface between linguistics and statistical-physics. More details about this project can be found at the URL `http://www.ime.usp.br/~tycho`.

3 A Gibbsian model for the interface between syntax and prosody

The model describes the way a speaker chooses a sentence in a given discursive context. This choice is made among all the sentences made available by his grammar. The basic idea is that this choice is driven by euphonic considerations. In other terms, the better the prosodic contour of a sentence fits the prosodic pattern of the language, the more likely this sentence is the one to be chosen.

A Gibbs state expresses this in a natural way. On one hand, each specific grammar defines a set of possible configurations of the system. On the other hand, prosody is described by the thermodynamical potential, which tells which configurations are more likely to occur, among the possible ones.

In a realistic linguistic situation, the description of the set of all possible configurations usually demands a detailed discussion of the values assumed by the relevant linguistic parameters. In order to simplify our presentation, we shall discuss here a simplified model, using what in Chomsky's hierarchy is called *regular grammars* (cf.[7,8]). This is a particular case of the model introduced in [10], to which we refer the reader for a more general presentation of our results, with rigorous mathematical proofs.

In this simplified model we shall suppose that the different grammars act on the same finite set of *word categories* Λ. From now on, we shall use the shorthand *word* instead of word category. The class of grammars we consider are topologically markovian and each grammar g can be identified by a matrix indexed by Λ and with entries equal to 0 or 1. The set of all the sentences of length n generated by g is the set

$$L_n(g) = \{(x_1, \cdots, x_n),\ x_j \in \Lambda,\ g(x_j, x_{j+1}) = 1,\ 1 \leq j \leq n-1\}\,. \quad (1)$$

This means that the grammaticality of a sentence $(x_1, \cdots, x_n)$, according to grammar g, depends only on the allowance of each transition (x_j, x_{j+1}) expressed by the fact that $g(x_j, x_{j+1}) = 1$. We will denote by $\mathcal{G}$ the set of all such grammars. It is natural to assume that the grammars in $\mathcal{G}$ are irreducible and aperiodic matrices.

From now on we shall use x_1^n as a shorthand for $(x_1, \cdots, x_n)$.

Let us now suppose that φ is a real function on Λ^m where m is a positive integer. For any $n > m$ we define a Gibbs state $\mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}}$ on $L_n(g)$ in the usual way by

$$\mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}}(x_1^n) = \frac{\exp[-H_\varphi(x_1^n)] \prod_{j=1}^{n-1} g(x_j, x_{j+1})}{Z_n(\varphi, g)}\,, \quad (2)$$

where

$$H_\varphi(x_1^n) = - \sum_{i=1}^{n-m+1} \varphi(x_i, \cdots, x_{i+m-1}) \tag{3}$$

and

$$Z_n(\varphi, g) = \sum_{y_1^n} \left\{ \exp\left[-H_\varphi(y_1^n)\right] \prod_{j=1}^{n-1} g(y_j, y_{j+1}) \right\} \tag{4}$$

is the partition function.

With this simple model it is possible to give a precise formulation to the question of how prosody provides hints to language acquisition. We are interested in the problem of identifying a grammar in $\mathcal{G}$ given a sentence x_1^n produced by a fixed but unknown grammar. We assume that the prosody, given by the potential φ, has been already acquired and is, therefore, known and fixed.

A first natural trial is to treat this as the statistical problem of estimating the grammar by a *Maximum Likelihood* procedure. This amounts to looking for the grammars g which maximize the probability $\mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}}(x_1^n)$. Since in this simple model the hamiltonian H_φ does not depend on g, the maximum likelihood procedure amounts to minimizing the partition function $Z_n(\varphi, g)$. Needless to say that this is a standard statistical physics question.

Let $\hat{g}_n(x_1^n)$ be the matrix defined by assigning the value 1 to all the entries which appear as transitions in the sentence x_1^n, and the value 0 to all the other entries. We now have the following theorem

Theorem 1 *For any φ and any $g_0 \in \mathcal{G}$, the estimator $\hat{g}_n(x_1^n)$ is the maximum likelihood estimator of the parental grammar, given the sample x_1^n. Moreover, there exists $\rho \in (0,1)$ such that for any n large enough*

$$\mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}_0}\left\{x_1^n : \hat{g}_n(x_1^n) = g_0\right\} \geq 1 - \rho^n \; . \tag{5}$$

For the proof we refer the reader to [10].

This result is quite satisfactory if all we need is a statement about the robustness of the acquisition procedure. It says that the learning child always identifies the parental grammar g_0 if he has a large enough sample of positive evidence. However, this is not what happens in real life, in which from time to time a generation of children chooses a grammar which is different from the parental one.

To improve the model in order to cover situations of language change we have two possibilities. The more realistic one is to add an additional feature to the model, namely the fact that a sentence is not only an ordered string of words, but it has also a syntactic structure produced by the grammar. This structure is not explicit in the sample and must be guessed by the learning child. In this extended model the maximum likelihood procedure may lead to the choice of a new grammar. An example of this will be presented in the next section.

The other possibility it is to use a different criterion to choose a grammar. Instead of minimizing the partition function $Z_n(\varphi, g)$, we shall look for a grammar g which minimizes the entropy of $\mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}}$.

Let us define the *entropy* of the Gibbs state $\mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}}$ defined on $L_n(g)$ as

$$h(\mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}}) = -\sum_{x_1^n} \mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}}(x_1^n) \log \mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}}(x_1^n) \ . \tag{6}$$

Given an ordered string of words x_1^n we define the Minimum Entropy Subset $\mathcal{E}_\varphi(x_1^n)$ by

$$\mathcal{E}_\varphi(x_1^n) = \{g \in \mathcal{G} \ : \ x_1^n \in L_n(g) \text{ and } \ h(\mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}}) \text{ is minimal } \} \ . \tag{7}$$

We may now introduce the *Minimum Entropy* procedure. Given φ, x_1^n the learning child chooses a grammar belonging to $\mathcal{E}_\varphi(x_1^n)$.

Let us define the variation $\mathbf{var}(\varphi)$ as

$$\mathbf{var}(\varphi) = \sup\{|\varphi(x_1^m) - \varphi(y_1^m)| \ : \ x_1^m \in \Lambda^m \ , y_1^m \in \Lambda^m\} \ . \tag{8}$$

Then the following theorem holds

Theorem 2 *There exists a positive real number r, such that for any potential φ such that $\mathbf{var}(\varphi) \leq r$ and any grammar g*

$$\lim_{n\to+\infty} \mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}} \left\{x_1^n \ : \ \mathcal{E}_\varphi^n(x_1^n) = \{g\}\right\} = 1 \ . \tag{9}$$

Theorem 2 says that the minimum entropy procedure coincides with the maximum likelihood procedure and identifies correctly the parental grammar, whenever the prosody is not too biased. This accounts for the robustness of the acquisition procedure. However, it is possible to choose the potential φ in such a way that the minimum entropy procedure leads to a new grammar. It is important to emphasize that this new grammar may be strictly greater than the maternal one. Here we are defining an order relation on $\mathcal{G}$ in the usual way: $g < g'$ means that $g(x,y) \leq g'(x,y)$ for all pair of words x and y and there exists at least one pair $(\bar{x},\bar{y})$ for which $g(\bar{x},\bar{y}) < g'(\bar{x},\bar{y})$.

This is the content of the next theorem.

Theorem 3 *For any g and g' in $\mathcal{G}$, such that $g < g'$, there exists a potential φ such that*

$$\lim_{n\to+\infty} \mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}} \left\{ x_1^n \ : \ g \notin \mathcal{E}_\varphi^n(x_1^n), g' \in \mathcal{E}_\varphi^n(x_1^n) \right\} = 1 \ . \tag{10}$$

We refer the reader to [10] for the proofs of theorems 2 and 3.

A simple example may help the reader to understand what is the situation expressed by theorem 3. Let us take $\Lambda = \{1,2\}$ and

$$g(1,1) = g(1,2) = g(2,1) = 1 \ \text{ and } g(2,2) = 0 \ . \tag{11}$$

Let's now take the potential φ acting only on pair of points ($m = 2$). If φ gives an overwhelming weight to the transition $(2,2)$ which is not allowed by g, then the minimum entropy procedure leads to the choice of the grammar g' allowing all the transitions ($g'(x,y) = 1$, for all pair of words x and y). In effect, g' is able to generate any sentence generated by g and moreover as sentences generated by $\mathbf{P}_{\mathrm{n}}^{\varphi,\mathrm{g}'}$ typically consist of very long sequences of 2, its entropy is very small.

Statistical analyses based on *entropy* considerations go back at least to the seminal work of Kullback [18], who showed that the notion of *relative entropy* appears naturally in maximal likelihood estimation. However, in our approach, entropy appears in a different way, close to the concept of *measure of diversity* like the *Shannon index* and *Rényi's α-entropy* (cf. [22]). A very nice related paper in the context dynamical systems is [6].

4 Structure identification and language change

A sentence produced by a grammar is not only an ordered string of words, it also has a structure. This structure is only indirectly indicated, through intonation, stress and other prosodic features. A native speaker is able to parse a sentence produced by his grammar, using these prosodic hints, as well as his own knowledge of the grammar. But a learning child, before he sets the parameters of Universal Grammar, must guess the structures of the sentences he receives. In this section we shall discuss this issue, by extending the model introduced in section 3.

In this extended model, a sentence will be an ordered string

$$(x_1, B_1, x_2, B_2, \cdots, B_{n-1}, x_n) \,, \tag{12}$$

where $x_1, \ldots, x_n$ are words belonging to Λ and $B_1, \ldots, B_{n-1}$ are hidden syntactical positions, which may either be occupied by a boundary mark $\mid$ or empty.

Let us call $\bar{\Lambda}$ the set

$$\bar{\Lambda} = \Lambda \cup \{\,|\,\} \,. \tag{13}$$

As before a grammar will be an element of $\bar{\mathcal{G}}$, the set of all matrices indexed by $\bar{\Lambda}$, with entries equal to 0 or 1. As in section 3, we shall assume that these matrices are irreducible and aperiodic. To avoid ambiguities, we shall also impose an extra constraint: for any $g \in \bar{\mathcal{G}}$ and any ordered couple of words (x, y), we have

$$g(x, y) g(x, \,|\,) g(\,|\,, y) = 0 \,. \tag{14}$$

Given a grammar $g \in \bar{\mathcal{G}}$, a sentence of length n generated by g is any ordered string $(x_1, B_1, x_2, B_2, \cdots, B_{n-1}, x_n, B_n)$, such that

$$\prod_{i=1}^{n-1} \chi_g(x_i, B_{i+1}, x_{i+1}) = 1 \,, \tag{15}$$

where

$$\chi_g(x_i, B_{i+1}, x_{i+1}) = g(x_j, x_{j+1}) \,, \text{ if } \; B_{j+1} \; \text{ is empty,} \tag{16}$$

and

$$\chi_g(x_i, B_{i+1}, x_{i+1}) = g(x_j, \,|\,) g(\,|\,, x_{j+1}) \,, \text{ if } \; B_{j+1} = | \;\, . \tag{17}$$

Before introducing the prosodic potential, we must add an extra detail to the picture. Words came from Λ with a stress mark. To simplify, we may assume that this mark has only two values, say *"stressed"* and *"unstressed"*, which will be represented by the symbols $+$ and $-$, respectively. The Boltzmann-Gibbs weight of a sentence will be a function only of the ordered string of stress marks of the words and the boundary marks $\mid$ which are present in the sentence.

Let φ be a real function on $\{-,+\}^2$. The hamiltonian $\bar{H}_\varphi$ will be defined as follows

$$\bar{H}_\varphi(x_1, B_1, x_2, \cdots, B_{n-1}, x_n) = \sum_{i=1}^{n-1} U_\varphi(x_i, B_{i+1}, x_{i+1}) \tag{18}$$

with

$$U_\varphi(x_i, B_{i+1}, x_{i+1}) \;=\; \begin{cases} 0 & \text{if } B_{i+1} = | \\ \varphi(s_i, s_{i+1}) & \text{if } B_{i+1} \text{ is empty} \end{cases} \tag{19}$$

where s_i and s_{i+1} are the stress marks of x_i and x_{i+1} respectively.

Now we define the Gibbs state $\mathbf{P}_{\mathrm{n}}^{\bar{\varphi},\mathrm{g}}$ as

$$\mathbf{P}_{\mathrm{n}}^{\bar{\varphi},\mathrm{g}}(x_1, B_1, \cdots, B_{n-1}, x_n) = \tag{20}$$

$$\frac{\exp[-\bar{H}_\varphi(x_1, B_1, \cdots, B_{n-1}, x_n)] \prod_{i=1}^{n-1} \chi_g(x_i, B_{i+1}, x_{i+1})}{\bar{Z}_n(\varphi, g)}\,, \tag{21}$$

where $\bar{Z}_n(\varphi, g)$ is the partition function.

A mother speaking to her child offers him a long sentence $(x_1, B_1, \cdots, B_{n-1}, x_n)$. But the only explicit data he receives is $(x_1, \cdots, x_n)$. He must estimate the hidden structure $(B_1, \cdots, B_{n-1})$, using his previous knowledge of φ. The estimation can be done using a *Maximum Likelihood* procedure. This amounts to look for a grammar g which assigns to that specific ordered string of words a sequence of syntactic marks $(B_1^\star, \cdots, B_{n-1}^\star)$ such that $\bar{H}_\varphi(x_1, B_1^\star, \cdots, B_{n-1}^\star, x_n)$ is minimum.

Remember that the mother must obey the interdictions of her own grammar, but the child is free to choose any grammar which is able to produce the string of words his mother offered to him. This opens the possibility of language change, even using the Maximum Likelihood procedure.

An example will help understanding this point. Let's take $\Lambda = \{1, 2\}$ and let's assume that the stress mark of 1 is $+$ and the stress mark of 2 is $-$. Let us suppose that the the maternal grammar is g defined as follows

$$g(1,1) = g(1,2) = g(2,1) = g(2,\ |\ ) = g(\ |\ ,2) = 1\ . \tag{22}$$

If $\varphi(1,\ |\ ) > 0$, $\varphi(1,\ |\ ) > 0$ and $\varphi(2,2) > 0$, the maximum likelihood procedure will lead to the choice of the grammar g' defined as follows

$$g'(1, \,|\,) = g'(\,|\,, 1) = g'(2, 2) = g'(1, 2) = g'(2, 1) = 1 \; . \qquad (23)$$

This example mimics in a very simplified way the mechanism behind the change which leads from Classical to Modern European Portuguese. The interested reader can find a detailed discussion of this in [13].

5 A simulated annealing model of language acquisition

The *Trigger Learning Algorithm* was recently introduced in the linguistic literature (cf.[15,11,1]). This algorithm models language acquisition by a child as a stochastic process taking values in the set $\mathcal{G}$ of all natural grammars. The algorithm explores $\mathcal{G}$ in a random way, deciding at each step of the procedure either to stay at the grammar at which it arrived at the previous step or to jump to a neighbor obtained by modifying the value of one randomly chosen parameter. This decision is taken under the stimulus of a random sample of sentences belonging to the parental language. The jump takes place if and only if this sample can be generated by the new but not by the actual grammar.

It follows from this definition that the algorithm stops its search when, for the first time, none of the grammars in the neighborhood is able to do any better than the actual grammar. This happens in particular (but not only) any time the parental grammar is reached.

Even if this issue is not raised in an explicit form in the linguistic literature, the aim of the algorithm is to achieve a maximization procedure, namely to find the grammars which maximize a given family of *evaluation* functions. However, given the way the algorithm is defined, the process may get trapped by grammars which are not global maxima of the family of fitness functions. This is a major problem for the Trigger Learning Algorithm since the fact of getting trapped somewhere is the only way it has to choose a grammar.

To avoid the problem of the algorithm getting trapped in grammars which are not global maxima there are two possibilities. The first one is to allow the algorithm to perform jumps between grammars which are not neighbors. This hypothesis does not seem to correspond to a realistic situation.

The second solution is obtained by weakening the restriction about jumps which do not increase the evaluation function, which will be no more forbidden, but only strongly depressed. In this section we show how to implement this alternative solution, by using a suitable generalization of the *simulated* *annealing* process (cf. [14] for a nice introduction to the subject, with applications).

Let us suppose that grammars are characterized by N binary parameters and that the set $\mathcal{G}$ of all possible grammars can be identified with the set $\{0,1\}^N$ of all ordered sequences of N elements assuming the values 0 and 1.

Two grammars g and g' are said to be *neighbors* if they have all the parameters set at the same values with the exception of one. Given $g \in \mathcal{G}$ and $i \in \{1, 2, \cdots, N\}$, let us denote by g^i the grammar obtained by setting all the parameters as in g, with the exception of the parameter of index i.

Let g_0 be the parental grammar. Let us call $L(g_0)$ the set of all the sentences offered as evidence of g_0 during acquisition. It is natural to assume that there exists a maximal length M for the sentences offered to the learning child during acquisition. Therefore we may take

$$L(g_0) = \cup_{n=1}^{M} L_n(g_0) \ , \tag{24}$$

where $L_n(g_0)$ denotes the set of all the sentences of length n generated by g_0.

Let us suppose that for each sentence η from the parental language $L(g_0)$, there exists an evaluation function f_η which associates a strictly positive real number to each grammar $g \in \mathcal{G}$.

A natural class of evaluation functions is the following. For any fixed grammar $g \in \mathcal{G}$, let w_g be the function which associates to each ordered string of words x_1^n either its structure according to g, in case there is one available , or a special symbol $\dagger$, in case there is none.

A mother offers a sentence $\eta = (x_1^n, w_{g_0}(x))$ from $L(g_0)$ to her child. However, the only explicit data received by the learning child is the string of words x_1^n. Let φ be the potential defining the parental prosody. Since the learning child has already acquired φ, he can evaluate a candidate grammar g looking at the value of $\bar{H}_\varphi(x_1^n, w_g(x))$. Therefore, it is natural to define f_η as the Boltzmann-Gibbs weight

$$f_\eta = \exp[-\bar{H}_\varphi(x_1^n, w_g(x))] \ , \tag{25}$$

with the convention that

$$\exp[-\bar{H}_\varphi(x_1^n, \dagger)] = 0 \ . \tag{26}$$

What follows does not depend on the particular way we define the class of evaluation functions.

The model is defined as follows. The mother offers a sequence of sentences to her child. Let us suppose that this choice is made at each step independently of the former choices and with the same law p. The distribution remains fixed during the evolution of the process. The only assumption we make about p it is that $p(\eta) > 0$ for every $\eta \in L(g_0)$.

Let us suppose that after $t-1$ steps, the algorithm has reached the grammar g. The choice of the position at time t id made as follows.

A parameter index $i \in \{1, \cdots, N\}$ is chosen at random and with uniform distribution. At time $t+1$ the process updates its value to g^i with probability

$$q_t(g, g^i) = \left(1 + \frac{f_{\eta_t}(g)}{f_{\eta_t}(g^i)}\right)^{\beta_t} , \tag{27}$$

where η_t is the sentence offered to the learning child at time t and $(\beta_1, \beta_2, \cdots)$ is a sequence of positive real numbers diverging sufficiently slowly to $+\infty$.

Let us call $\{G_t, t = 0, 1, \cdots\}$ the non homogeneous Markov chain defined this way. By construction its transition probability matrix at time t is

$$Q_t(g, g^i) = \frac{1}{N} \sum_{\eta \in L(g_0)} p(\eta) q_t(g, g^i) , \tag{28}$$

for any g and any i and

$$Q_t(g, g) = 1 - \sum_{i=1}^{N} Q_t(g, g^i) . \tag{29}$$

Let us call π_t be its invariant probability measure.

The existence of the limits

$$\lim_{t \to +\infty} \pi_t(g) = \pi(g) , \tag{30}$$

and

$$\lim_{t \to +\infty} \mathbf{P}\{G_t = g\} = \pi(g) , \tag{31}$$

for any $g \in \mathcal{G}$, follows in a standard way under conditions like

$$\beta_t \le C \log t , \tag{32}$$

where $C > 0$ is a suitable constant which depends only on the family of evaluation functions (cf. [16,17]).

The interesting issue here is to find necessary and sufficient conditions on the family $\{f_\eta, \eta \in L(g_0)\}$ assuring that the limit distribution π is a Dirac measure. In effect, it is reasonable to expect that in a given community, in which all the adults have the same grammar and prosody, all the learning children will converge to the same and unique grammar. It is important to emphasize that this unique grammar may be different from the grammar spoken by the generation of the adults.

The following theorem holds

Theorem 4 *A sufficient condition for the process (G_t) to converge in law to a Dirac measure is that there exists a grammar $\bar{g}$, a sentence $\bar{\eta}$ and an index $\bar{i}$ such that*

$$\frac{f_{\bar{\eta}}(\bar{g})}{f_{\bar{\eta}}(\bar{g}^{\bar{i}})} > \frac{f_\zeta(g)}{f_\zeta(g^j)}\,, \tag{33}$$

for all $g \neq \bar{g}$, and all j and ζ.

This section is based on [2–4]. It will appear in a more general context in a forthcoming article [5].

Acknowledgements

This article is an extended version of the talk presented at StatPhys20 by one of the authors, A.G. We thank the audience of the talk, in particular Michael Fisher and Joel Lebowitz, for interesting remarks and questions. We are also indebted to Maria Bernadete Abaurre, Robert Berwick, Robert Frank, Elisabeti Kira, Anthony Kroch and Artur Lopes for many illuminating discussions these last years.

References

[1] R. Berwick and P. Niyogy. Formalizing triggers: a learning model for finite space, *Linguistic Inquiry* **27**, 605-622, 1997.

[2] M. Cassandro and A. Galves. Acquisition et changement linguistique dans le modèle de Gibson et Wexler, *Colloque Langues et Grammaire 2*, Université de Paris VII, June 8-10, 1995.

[3] M. Cassandro and A. Galves. Language acquisition and change in a generalized Gibson-Wexler model, *Fourth Meeting on Mathematics of the Language (MOL4)* , University of Pennsylvania, Philadelphia, October 27-28, 1995. Tarragona, Spain, May 2-4, 1996.

[4] M. Cassandro, A. Galves and C. Galves. Structure recognition and language change in a generalized GW model, *II International Conference on Mathematical Linguistics (ICML'96)*,

[5] M. Cassandro, A. Galves and C. Galves. Structure identification and language change in a thermalized GW model. *Work in progress.*

[6] J.-R. Chazottes, E. Floriani and R. Lima. Relative entropy and identification of Gibbs measures in dynamical systems *J. Statist. Phys.* **90**, 697–725, 1998.

[7] N. Chomsky. Three models for the description of language. *IRE Trans. on Inform. Theory*, **IT 2**, 113-124, 1953.

[8] N. Chomsky. Formal properties of grammars, in: *Handbook of Math. Psych.*, **2**, 323-418, John Wiley, New York, 1963.

[9] N. Chomsky. *A minimalist program for linguistic theory* The MIT Press, Cambridge, MA, 1995.

[10] P. Collet, A. Galves and A. Lopes. Maximum likelihood and minimum entropy identification of grammars, *Random and Computational Dynamics* **3**, 241-250, 1995.

[11] R. Frank and S. Kapur. On the use of triggers on parameter setting, *Linguistic Inquiry* **27**, 623-660, 1997.

[12] C. Galves, Clitic Placement and Parametric Changes in Portuguese, in: *Aspects of romance linguistics, Selected papers from the Linguistic Symposium on Romance languages XXIV*, Georgetown University Press, 1996.

[13] A. Galves and C. Galves. A case study of prosody driven language change. Preprint (can be retrieved at URL `http://www.ime.usp.br/~tycho`).

[14] D. Geman. Random fields and inverse problem in imaging, in: *18e Ecole d' été de probabilités de St Flour*, Lecture Notes in Mathematics, Berlin, Heidelberg, New York: Springer, 1990.

[15] E. Gibson and K. Wexler. Triggers, *Linguistic Inquiry* **25**, 407-454, 1994.

[16] B. Gidas. *Metropolis-type Monte Carlo simulation algorithms and simulated annealing*, Brown University, 1991.

[17] B. Hajek. Cooling schedules for optimal annealing, *Math. Oper. Research* **13**, 311-329, 1988.

[18] S. Kullback. *Information theory and statistics* John Wiley, New York, 1959.

[19] G. Marcus. Negative evidence in language acquisition, *Cognition* **46**, 1993.

[20] J. Morgan. *From Simple Input to Complex Grammar* The MIT Press, Cambridge, MA, 1986.

[21] M. Nespor, M.T. Guasti and A. Christophe. What can infants learn from prosodic constituents? *18th GLOW Colloquium*, Tromso, 1995.

[22] A. Rény. On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 547-561, 1961.