IDENTIFYING FEATURES IN THE PRESENCE OF COMPETING EVIDENCE. THE CASE OF FIRST-LANGUAGE ACQUISITION

ROBERTO FERNÁNDEZ
Instituto de Estudos Avançados, Universidade de São Paulo,
Av. Prof. Luciano Gualberto, Travessa J, 374 térreo, Cidade Universitária,
05508-900 São Paulo, SP, Brazil,
rf@ime.usp.br

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

We discuss mathematical issues suggested by the processes of first-language acquisiton and language change. We present a model of language acquisition with two components: A probability measure describing sentence selection by a native speaker, and an identification principle modeling how the child choses an element of the finite set of natural grammars. More generally, we present an approach to the problem of classifying existing evidence to choose among a finite set of policies in the presence of possibly conflicting hints.

1 Introduction

The *Principles-and-Parameters* approach to the theory of grammar developed by Chomsky and collaborators in the last decades claims 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 —the *Universal Grammar*. The learning child identifies a grammar by assigning structures to a sample of utterances from the parental language.

To make guesses about the structures of the sentences the child must compare possibly competing evidence provided, in particular, by the parental prosody[26,27]. This process does not always lead to the parental grammar. The failure to do so is what is called *language change*.

In the present paper we discuss mathematical issues which are expected to be relevant to these issues. Language acquisition is our main motivation and laboratory and, in fact, the results reported here are part of a long-term study on the subject[5]. Nevertheless, our approach is potentially more general and refers to the problem of classifying existing evidence to choose among a finite set of policies in the presence of possibly conflicting hints. Less compromisingly, the reader can consider our framework simply as a (non-simple) mathematical problem, and reduce linguistics to a convenient, and inspiring, source of questions and nomenclature.

2 A mathematical model for the interface between syntax and prosody

Our discussion is centered on a recently introduced model of language acquisition[20,9,14,20,6]. The model has two components: The first one, discussed in the present section, is a probability measure describing the process of sentence selection by a native speaker. The second component, presented in the next two sections, is an identification principle modeling how the child choses an element of the *finite* set of natural grammars.

The probability measure must take into account both syntax and prosody. The *syntax* is the computational system of the grammar. This is a finite set of constraints, of algebraic nature, which determine, in a categorical way, which sentences are possible. For brevity we will say *grammar* instead of *computational system of the grammar. Prosody* maps sentences into acoustic realizations. It also imposes constraints albeit of a softer nature than grammar.

We put prosody and syntax together through a Gibbsian distribution. Each specific grammar sets the possible configurations of the system, while prosody is described by a thermodynamical potential that favors some possible configurations (sentences) over others.

Informally speaking, a sentence produced by a grammar is composed by an ordered string of words together with a structural description (from now on we use the shorthand *structure*). Let us call $\mathcal{X}$ the finite set of *words* and $\mathcal{B}$ the finite set of *structural symbols.* Sentences are ordered strings of elements of $\Lambda = \mathcal{X} \cup \mathcal{B}$. A grammar is characterized by a finite set $\Delta \subset \mathbf{Z}$ —the *control range*— and a set of allowed strings $G \subset \Lambda^{\Delta}$. We fix Δ once and for all and we identify each grammar with its corresponding set G. The *language* generated by G is the set

$$L(G) \ = \ \Big\{ s = (s_n)_{n\in\mathbf{Z}} \in \Lambda^{\mathbf{Z}} : \ (s_{n+m})_{m\in\Delta} \in G, \text{ for any } \ n \in \mathbf{Z} \Big\} \, . \tag{1}$$

The prosody will be defined through a Hölder continuous real function φ on $\Lambda^{\mathbf{N}}$, which will be called the *potential.* In the linguistic case, it is reasonable to assume that φ actually depends only on a finite set of coordinates. The probability measure $\mathbf{P}^{\varphi,\mathrm{G}}$ is defined as the unique measure on $\Lambda^{\mathbf{Z}}$ with the property that there is a positive constant $C > 1$, such that for any element t

of $L(G)$ and for any positive integer N we have

$$\frac{1}{C} \le \frac{\mathbf{P}^{\varphi,\mathrm{G}}(t_1^N)}{e^{-NP+\sum_{j=0}^{N-1}\varphi(\sigma^j(t))}} \le C\ , \tag{2}$$

where $P = P(\varphi, G)$ is the pressure associated to the potential φ on $L(G)$, σ is the usual shift on $\Lambda^{\mathbf{Z}}$ and

$$t_1^N = \{s = (s_n)_n \in L(G) : \ s_n = t_n \ , \text{ for } \ n = 1 \cdots N\}\ . \tag{3}$$

The classical references on this view of Gibbsian measures is the book by Bowen[3] to which we refer the reader. A more recent reference is the extensive and up-to date review by Parry and Pollicott[28].

The simplest situation is when $\mathcal{X} = \{0, 1\}$ and $\mathcal{B} = \emptyset$. These are called *regular grammars* in Chomsky hierarchy. While they are knowon to be too simple to be realistic regarding human languages, they exhibit some features of the general phenomena. See the original articles of Chomsky[12,13] for criticism and general presentation of formal languages. A regular grammar G is therefore entirely specified by a matrix —which we denote also as G— indexed by $\mathcal{X}$ and with entries equal to 0 or 1. The language generated by G is the set

$$L(G) = \{(x_n), \in \mathcal{X}^{\mathbf{Z}} : G(x_n, x_{n+1}) = 1, \ \text{ for any } \ n \in \mathbf{Z}\}\ . \tag{4}$$

In dynamical systems this is called a subshift of finite type. If φ depends only on the first two coordinates $\mathbf{P}^{\varphi,\mathrm{G}}$ is the law of a Markov chain.

3 Language acquisition and change

We focus on the following mathematical problem. Suppose we know the potential φ but ignore the grammar. Can we identify it if we are given a (large) sample of the language, that is a sentence x_1^n produced with distribution $\mathbf{P}^{\varphi,\mathrm{G}}$? We interprete this question as a model of how a known prosody provides the child with hints for language acquisition. The unknown G represents the "parental grammar".

In this section we review results obtained by Collet, Galves and Lopes[14], who addressed this issue in the framework of regular grammars (see also Chazottes, Floriani and Lima[11] for related work in a non-linguistic context). The following theorem gives the *maximum likelihood* estimator of such grammars.

Theorem 1 *For any φ and any regular $G_0 \in \mathcal{G}$, the maximum likelihood estimator of the parental grammar, given the sample x_1^n, is the matrix $\widehat{G}_n(x_1^n)$ 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. Moreover, there exists $\rho \in (0,1)$ such that for any n large enough*

$$\mathbf{P}^{\varphi,\mathrm{G}_0}\left\{x_1^\infty : \hat{G}_n(x_1^n) = G_0\right\} \geq 1 - \rho^n \ . \qquad (5)$$

In words, an identification procedure based in the maximum likelihood estimator is robust. The learning child always identifies the parental grammar G_0 once s/he listens to a large enough sample. We are interested, however, in procedures that allow a certain amount of miss-identification, to model the observed fact that from time to time a generation of children chooses a grammar different from the parental one. A possible approach is provided by the following *minimum entropy principle.* Given a string s, we define the *minimum entropy subset* corresponding to a sample of size n as

$$\mathcal{E}_\varphi^n(s) = \left\{G \in \mathcal{G}\ s_1^n \subset L(G) \quad \text{and} \quad h(\mathbf{P}^{\varphi,\mathrm{G}}) \quad \text{is minimal}\right\} \ , \qquad (6)$$

where $h(\mathbf{P}^{\varphi,\mathrm{G}})$ is the Kolmogorov-Sinai entropy of $\mathbf{P}^{\varphi,\mathrm{G}}$ (we refer the reader to Bowen's book[3], for instance, for the definition of entropy). The *minimum entropy identification procedure* says that the learner chooses a grammar belonging to $\mathcal{E}_\phi^n(s)$.

While relative entropy appears naturally in maximal likelihood estimations[24], in the present approach it is used more as a *measure of diversity* like the *Shannon index* and *Rényi's α-entropy.* The following two theorems show that this procedure can be interpreted as a generalization of the maximum-likelihood procedure allowing grammar change.

The first theorem says that both the maximum likelihood and minimum entropy procedures agree as long as the prosody is not too biased.

Theorem 2 *There exists a neighborhood $\mathcal{O}$ of the constant function, such that for any φ in $\mathcal{O}$ and any regular grammar G_0 the minimum entropy sets $\mathcal{E}_\varphi^n(s)$ converge to G_0 for $\mathbf{P}^{\varphi,\mathrm{G}_0}$-almost-all choices of s, as n diverges.*

The second theorem shows that grammar change can be driven by a biased prosodic potential.. We endow $\mathcal{G}$ with the natural partial order: $G < G'$ if $G(x,y) \leq G'(x,y)$ for all pairs of words (x,y) and there exists at least one pair $(\bar{x},\bar{y})$ for which $G(\bar{x},\bar{y}) < G'(\bar{x},\bar{y})$.

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

$$\lim_{n\to+\infty} \mathbf{P}^{\varphi,\mathrm{G}}\left\{s \ : \ G \notin \mathcal{E}_\varphi^n(s)\, ,\, G' \in \mathcal{E}_\varphi^n(s)\right\} = 1 \ . \qquad (7)$$

4 Structure recognition and language acquisition

Sentences produced by general grammars are not only ordered strings of words but also have structure. In normal speech, however, the information regarding structure is given indirectly through prosodic features such as intonation, stress, etc. A native speaker is able to use these prosodic hints and his knowledge of the grammar to parse a sentence. But a learning child that has not yet set the parameters of the Universal Grammar is in fact forced to guess the structures of the sentences s/he receives. This guess will influentiate the estimation of the underlying grammar possibly leading to language change.

Formally, a mother/father offers the child a long sentence, which is an ordered string $(t_1, \cdots, t_k)$. Some of the t_i are words belonging to $\mathcal{X}$ and some are structural symbols —that is, elements of $\mathcal{B}$— representing structural descriptions. For notational simplicity let us assume that they alternate, that is the string is of the form

$$(x_1, B_1, x_2, B_2, \cdots, B_{n-1}, x_n) \ , \tag{8}$$

where $x_i \in \mathcal{X}$ and $B_j \in \mathcal{B}$ (this is no loss of generality, as we assume that there is a possible value "$\emptyset$" both in $\mathcal{X}$ and $\mathcal{B}$ which indicates "absence of" word or structural symbols at the given position).

The only explicit data the child receives is $(x_1, \cdots, x_n)$. S/He must estimate the hidden structure $(B_1, \cdots, B_{n-1})$, using his previous knowledge of φ and at the same time estimate the grammar. An estimation done through a maximum likelihood procedure yields a grammar which assigns to the string $(x_1, \cdots, x_n)$ a sequence of syntactic marks $(B_1^\star, \cdots, B_{n-1}^\star)$ such that

$$H_\varphi(x_1, B_1^\star, \cdots, B_{n-1}^\star, x_n) \ = \ \sum_{j=1}^{2n-m} \varphi\left(\sigma^j(x_1, B_1^\star, \cdots, B_{n-1}^\star, x_n)\right) \tag{9}$$

is minimum. Here m is the range of the prosodic potential φ, that is, the number of coordinates on which it depends and σ is the shift.

Let us illustrate this fact through an example. We take $\mathcal{X} = \{1, 2\}$ and $\mathcal{B} = \{\,|\,, \emptyset\}$ where "$|$" is interpreted as a boundary mark. Let us suppose that the the maternal grammar is G_0 defined by the assignments

$$G_0(1,1) = G_0(1,2) = G_0(2,1) = G_0(2,\ |\ ) = G_0(\ |\ ,2) = 1\ . \tag{10}$$

If the prosodic potential satisfies $\varphi(1,\ |\ ) > 0$, $\varphi(1,\ |\ ) > 0$ and $\varphi(2,2) > 0$, the maximum likelihood procedure will lead to a grammar $G\prime$ defined as follows

$$G(1,\ |\ ) = G(\ |\ ,1) = G(2,2) = G(1,2) = G(2,1) = 1\ . \tag{11}$$

This example shows a caricature of the mechanism behind the change from Classical to Modern European Portuguese[20].

5 Acquisition with maturation

We present now a *Maturation Model of Language Acquisition* (MMLA)[7,8,9] which is an extension of the *Trigger Learning Algorithm* (TLA)[21,19,2]. The latter is a stochastic process that explores $\mathcal{G}$ in a random way, deciding at each step whether to stay at the current grammar or to jump to a neighbor grammar 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. TLA stops its search when, for the first time, none of the grammars in the neighborhood is able to do better than the actual grammar. This happens, in particular, each time the parental grammar is reached.

In TLA the decision depends on a boolean *evaluation function* that says "go" in case of improvement. In MMLA more general evaluation functions are allowed to accomodate prosody. Furthermore, MMLA do not forbid jumps that do not increase the evaluation function; it only discourages them. This discouragement increases with time, mimicking the effect of *maturation* during acquisition.

Without loss of generality we consider grammars characterized by a finite number of binary parameters. Two grammars G and G' are said to be *neighbors* if they set all but one of the parameters at the same value.

Associated to each utterance $\underline{x}$ from the parental language there is an evaluation function $f_{\underline{x}} : \mathcal{G} \to]0, \infty[$. It is natural to define them through Boltzmann-Gibbs weights:

$$f_{\underline{x}} = \exp[-\bar{H}_\varphi(w_G(\underline{x}))] \ . \tag{12}$$

Here, w_G is the function that associates to each utterance a complete sentence, that is the string of words and its structure. In case there is no structure available for the utterance, the function w_G associates to it a special symbol $\dagger$. The exponent H_φ is a *cost function* defined analogously to (9). We use the convention that

$$\exp[-\bar{H}_\varphi(\dagger)] = 0 \ . \tag{13}$$

The evolution of MMLA is driven by a sequence $\underline{x}(\tau)$, $\tau = 1, 2, \ldots$ of utterances from the parental language. These utterances are chosen independently and with the same law. Let us suppose that after $\tau - 1$ steps, MMLA has reached grammar G. To determine its value at time τ, a candidate G' is chosen among the neighbors of G with uniform distribution. The process accepts it with probability

$$q_\tau(G, G') = \left(1 + \frac{f_{\underline{x}(\tau)}(G)}{f_{\underline{x}(\tau)}(G')}\right)^{-\beta_\tau} , \tag{14}$$

where $(\beta_1, \beta_2, \cdots)$ is a sequence of positive real numbers diverging sufficiently slowly to $+\infty$.

Let us call $\{\mathbf{G}_\tau, \tau = 0, 1, \cdots\}$ the non homogeneous Markov chain defined in this way. The existence of the limit

$$\lim_{\tau \to +\infty} \mathbf{P}\left\{\mathbf{G}_\tau = G\right\} = \pi(G) , \tag{15}$$

for any $G \in \mathcal{G}$, follows in a standard way under conditions like $\beta_\tau \leq C \log \tau$, where $C > 0$ is a suitable constant which depends only on the family of evaluation functions[22,23].

The interesting issue here is to find necessary and sufficient conditions on the evaluation functions assuring that the limit distribution π is a Dirac measure. In effect, it is reasonable to expect that if in a community all the adults have the same grammar and prosody, then all the learning children will also adopt a unique grammar. Nevertheless, historical examples show that the latter may differ from the grammar spoken by the adults[10].

6 Long-range costs and the presence of phase transitions

A natural extension of the previous model would correspond to letting the potential φ in (2) to depend on infintely many coordinates. In such a situation, the resulting model could exhibit coexistence of phases (associated to a first-order phase transition). Roughly speaking this means that the probability distributions for finite strings have more than one infinite-string limit.

Already the very definition of the model gets more complicated in this case. Indeed, in the presence of "long-range prosody" the whole of the (infinite) discourse has to be taken into account even when focusing at finite strings. We refer the reader to Ruelle[29] for the appropriate formalism.

One may wonder on the need for such generality for linguistic applications. One may imagine, for instance, that rhythm is produced by almost periodic stress patterns, indicating the presence of persistent long-range adjustments in the stress contours. The main new aspect brought by infinite-range interactions is the loss of uniqueness in the infinite-volume measure. Can this be related to language change?

Role of observables at infinity

The finite-volume distributions are not enough to determine the phase if there is more than one. Indeed, by definition, *all* phases have the *same* finite-volume conditional probabilities. The observations that univocally pinpoint each phase are associated to *observables at infinity*, that is to functions whose value is not altered by the change of finitely many words. In this situation, conclusions based on observing a finite "window" can easily be missleading. For instance, if a finite sample of the two-dimensional Ising model shows an overwhelming fraction of up spins, it is not easy to decide whether the sample comes from the "+" phase at zero field or from a system with a large magnetic field. An individual deciding prematurely can incorporate an inexisting field in the description. In practice, a large field plays the role of a hard constraint, and the wrong choice amounts, in the previous language, to a grammar change. We observe that such a change is not associated with the presence of undetected structural symbols.

Possible non-Gibbsianness

In the framework of Section 4, the child is not considering the original measure, but only its *projection* on the part of the configurations formed only by words. In statistical mechanics this is known as a *decimated measure*. It is a well known fact that, in the presence of phase transitions a decimated Gibbs measure may cease to be Gibbsian[17]. While this type of measures have been subjected to much study in the last decade[18,15], still the practical consequences of non-Gibbsianness remain to be clarified. Nevertheless, let us mention the consequences of two of the scenarios being observed. On the one hand, some of this non-Gibbsian measures can be "forced" into a generalized Gibbsian framework by restricting the configuration space[4]. This restrictions *depends on the phase* whose decimation is being observed. This admits an immediate interpretation in terms of grammar change. On the other hand, a further decimation may bring the measure back to honest Gibbsianness, but with a potential that, once again, *depends on the phase*[25]. If the couplings responsable for this difference are sufficiently strong, this selectivity is tantamount to a grammar change.

Sensitivity to the choice of the prosodic potential

It is simple to see that there is considerable freedom in the choice of the functions φ. The interaction can be "rewritten" in many different ways while leaving the conditional probabilities unchanged. In the presence of longrangeness, this freedom could make things more delicate. Several results, chiefly the smoothness of the resulting expectations —and in particular the presence of exotic phase transitions— crucially depend on the way interactions are written or, more formally, on the space of interactions being considered[16]. While this may sound too abstract for the present discussion, the moral is clear. Not all transcriptions of pratical requirements into prosodic potentials are equally performing. The problem of finding the "optimal" choice is difficult and probably model-dependent, but sooner or later it may have to be tackled. This remark is specially valid if long-range interactions are included.

Acknowledgments

This article is an extended version of the talk presented at the meeting *Dynamical Systems: from Cristal to Chaos* by one of the authors, A.G. We are indebted to M. B. Abaurre, R. Berwick, M. Cassandro, P. Collet, R. Frank, Ch. Galves, S. Frota, A. Kroch, E. Rodrigues and M. Vigário for many illuminating discussions these last years.

This work was done as part of FAPESP's Projeto Temático *Rhythmic patterns, parameter setting and language change* (grant 98/3382-0), and FINEP's Núcleo de Excelência *"Critical phenomena in probability and stochastic processes"*, (PRONEX grant 41.96.0923.00). The authors were partially supported by CNPq [grants 301625/95-6 (RF) and 301301/79 (AG)].

References

1. M. B. Abaurre and C. Galves. As diferencas rítmicas entre o português europeu e o português brasileiro: uma abordagem otimalista e minimalista. D.E.L.T.A. **14** no 2, 377-423, 1998.
2. R. Berwick and P. Niyogy. Formalizing triggers: a learning model for finite space, *Linguistic Inquiry* **27**, 605-622, 1997.
3. R. Bowen, *Equilibrium states of Anosov systems.* Lect. Notes in Math. **470**, Springer, New York, 1975.
4. J. Bricmont, A. Kupiainen and R. Lefevere. Renormalization group pathologies and the definition of Gibbs states, *Commun. Math. Phys.*, **194**:359–388, 1998.
5. Project *Rhythmic patterns, parameter setting and language change,* URL `http://www.ime.usp.br/~tycho`).
6. Marzio Cassandro, Pierre Collet, Antonio Galves, and Charlotte Galves. A Statistical Physics Approach to Language Acquisiton and Language Change. *Physica A*, 263:427–437, 1999

7. 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.
8. 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.
9. 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)*, 1996.
10. M. Cassandro, A. Galves and E. Rodrigues. A thermalized model of language acquisition, *Work in progress*, 1999.
11. 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.
12. N. Chomsky. Three models for the description of language. *IRE Trans. on Inform. Theory*, **IT 2**, 113-124, 1956.
13. N. Chomsky. Formal properties of grammars, in: *Handbook of Math. Psych.*, **2**, 323-418, John Wiley, New York, 1963.
14. P. Collet, A. Galves and A. Lopes. Maximum likelihood and minimum entropy identification of grammars, *Random and Computational Dynamics* **3**, 241-250, 1995.
15. A. C. D. van Enter. The renormalization-group peculiarities of Griffiths and Pearce: What have we learned?, Preprint `http://www.ma.utexas.edu/mp_arc-bin/mpa?yn=98-692`, 1998.
16. A. C. D. van Enter and R. Fernández. A remark on different norms and analyticity for many-particle interactions, *J. Stat. Phys.*, **56**, 965–972, 1989.
17. A. C. D. van Enter, R. Fernández and A. D. Sokal. Regularity properties and pathologies of position-space renormalization-group transformations: Scope and limitations of Gibbsian theory, *J. Stat. Phys.*, **75**, 879–1167, 1993.
18. R. Fernández. Random fields in lattices. The Gibbsianness issue, *Resenhas IME-USP*, **3**, 391–421, 1998.
19. R. Frank and S. Kapur. On the use of triggers on parameter setting, *Linguistic Inquiry* **27**, 623-660, 1997.
20. 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/papers/lang_change.ps`).
21. E. Gibson and K. Wexler. Triggers, *Linguistic Inquiry* **25**, 407-454, 1994.

22. B. Gidas. *Metropolis-type Monte Carlo simulation algorithms and simulated annealing*, Brown University, 1991.
23. B. Hajek. Cooling schedules for optimal annealing, *Math. Oper. Research* **13**, 311-329, 1988.
24. S. Kullback. *Information theory and statistics* John Wiley, New York, 1959.
25. J. Lörinczi and K. vande Velde. A note on the projection of Gibbs measures, *J. Stat. Phys.*, **77**, 881–887, 1994.
26. J. Morgan, *From Simple Imput to Complex Grammar*, MIT Press, Cambridge, MA, 1986.
27. M. Nespor, M.T. Guasti and A. Christophe. What can infants learn from prosodic constituents? *18th GLOW Colloquium*, Tromso, 1995.
28. W. Parry and M. Pollicott, *Zeta functions and the periodic orbits structure of hyperbolic dynamics.* Astérisque **187-188**, 1990.
29. D. Ruelle, *Thermodynamic Formalism*, Wesley, Reading, 1978 .