Oudeyer, P-Y (2005) How phonological structures can be culturally selected for learnability, Adaptive Behavior, 13(4), pp. 269--280.

Adaptive **Behavior**

How Phonological Structures Can Be Culturally Selected for Learnability

Pierre-Yves Oudeyer
Sony CSL Paris, France

This paper shows how phonological structures can be culturally selected so as to become learnable and adapted to the ecological niche formed by the brains and bodies of speakers. A computational model of the cultural formation of syllable systems illustrates how general learning and physical biases can influence the evolution of the structure of vocalization systems. We use the artificial life methodology of building a society of artificial agents, equipped with motor, perceptual and cognitive systems that are generic and have a realistic complexity. We demonstrate that agents, playing the "imitation game," build shared syllable systems and show how these syllable systems relate to existing human syllable systems. Detailed experiments study the learnability of the self-organized syllable systems. In particular, we reproduce the critical period effect and the artificial language learning effect without the need for innate biases which specify explicitly in advance the form of possible phonological structures. The ability of children agents to learn syllable systems is explained by the cultural evolutionary history of these syllable systems, which were selected for learnability.

Keywords origins of speech · evolution · acquisition · constraint · phonetics · phonology · learnability

1 The Principle of Cultural Selection for Learnability

Children learn language, and in particular sound systems, incredibly easily and quickly, in spite of its apparent idiosyncratic complexity and noisy learning conditions. Many researchers (Pinker & Bloom, 1990) believe this can not be possible without a substantial genetic preprogramming of specific neural circuits that encode explicitly at birth the main structures of language. In fact the role of learning in language development is even sometimes thought to be very minor (Piattelli-Palmarini, 1989) and reduced to the setting of a few parameters as in the Principles and Parameters theory (Chomsky & Halle, 1968) or in Optimality Theory (Archangeli & Langendoen, 1997). Yet, a growing number of researchers have challenged this view, and have argued that no linguistically specific innate neural device is necessary to account for the oddities of language learning (and structure): Rather, they propose that they result from the complex interactions between a number of general motor, perceptual, cognitive, social and functional constraints during the course of cultural interactions (Steels, 1997; Kirby, 2001; Tomasello, 2003). According to this view, language emerged and evolved so as to fit the ecological niche of initially non-speaking human brains and bodies: Languages were (and still are) culturally selected so as to be learnable.

As a consequence, if one wants to understand the principles of language acquisition and the structure of

Correspondence to: Pierre-Yves Oudeyer, Sony Computer Science Laboratory Paris, 6, rue Amyot, 75005 Paris, France.
E-mail: py@csl.sony.fr *Web*: http://www.csl.sony.fr/~py
Tel.: +33144080503, *Fax*: +33145878750.

(2005), Vol 13(4): 269–280.
[1059–7123(200512) 13:4; 269–280; 059568]
Figures 1, 2 appear in color online: http://adb.sagepub.com

language, one has to understand the principles of language emergence and evolution. Conversely, if one wants to understand the principles of language emergence and evolution, it is a necessity to understand the principles of language acquisition, since the biases due to the learning mechanism and the learning situations influence crucially the shaping of linguistic structures. There exists already a vast amount of computational models of the emergence and evolution of language (Steels, 1997; Kaplan, 2001; Kirby, 2001; Cangelosi & Parisi, 2002; de Boer, 2001; Oudeyer, 2005b; Vogt, 2003). Yet, those who have closely studied the impact of learning biases upon the evolution of linguistic structures, and in particular the cultural selection for better learnability, are still quite rare. Zuidema (2003) presented abstract simulations of the formation of syntactic structures and detailed the influence of cognitive constraints upon the generated syntax. Brighton et al. (2005) presented a thorough study of several simulations of the origins of syntax (Kirby, 2001) which were re-described in the light of this paradigm of cultural selection for learnability. The objective of this paper is to show an example of cultural selection for learnability in the field of phonology. We will present a computational system which illustrates how phonological structures can be culturally selected so as to become learnable.

More particularly, our system models the cultural formation of syllable systems, which are thought to be a fundamental unit of the complex phonological systems of human languages (MacNeilage, 1998). It relies on the interactional framework developed by de Boer (de Boer, 2001) and called the "imitation game." de Boer (2001) developed this framework in the context of simulations of the origins and evolution of vowel systems. He showed how self-organization may have allowed agents to form efficient vowel systems in terms of acoustic distinctiveness, and with neither explicit optimization nor centralized control. Lindblom (1992) had shown that one could predict the most frequent vowel systems in human languages by searching those that were acoustically maximally distinctive. The work of de Boer allows us to understand how this optimization can be actually realized implicitly and through cultural evolution in a society of agents.

We extend his simulations by providing the agent with the capacity to produce complex dynamic utterances, built by the sequencing of articulatory targets. This ability to sequence targets (or phonemes) allows us to study the evolution of phonological structures, i.e., the rules of phoneme sequencing. The learning of syllable repertoires is controlled by a mechanism that we will detail in order to understand its biases. Furthermore, we introduce the notion of articulatory/energetic cost for vocalizations, which impacts the learning system. This allows us to embody in an operational system the "ease versus distinctiveness" hypothesis proposed by Lindblom (1992). The introduction of an articulatory cost and its role in the shaping of syllable repertoires was also studied in the computational model presented by Redford et al. (2001). Yet, Redford et al. (2001) used explicit centralized optimization like Lindblom, and did not study how this could be realized in a society of decentralized autonomous agents. A previous paper (Oudeyer, 2001) studied the structure of the self-organized shared syllable repertoires that appear in our system, and presented comparisons with actual human syllable repertoires, as was done in Lindblom (1992), Redford et al. (2001), and de Boer (2001). The focus of this paper is different and new as compared to these previous studies: We will here concentrate on the study of the learnability of the self-organized syllable systems by children and adult agents, and compare it to the learnability of randomly generated artificial syllable systems.

In three other papers (Oudeyer, 2005a,b,), we developed another model of the origins of phonological structures that must not be confused with the one that we present in this paper. In both models, agents are able to produce complex utterances, and rules of sequential combination emerge. But the assumptions as well as the objective of this other model are very different. First of all, this other model does not assume an explicit pressure for building repertoires of distinctive sounds: Agents do not play the "imitation game," and there are no rewards or feedback between agents. In fact, agents have no skills to coordinate socially, and are just coupled by the fact that they can hear the vocalizations of the other agents with which they share the same environment. This means that the model did not pre-suppose any convention such as the rules of a language game. That model was designed to study the bootstrapping of speech: How did the first speech code appear when linguistic communication did not exist yet? In contrast, the model we present here, as well as the one of de Boer, deals with the cultural formation and evolution of a speech code once the ability to have linguistic interactions has appeared, with all the associated tools such as the interactional

Oudeyer, P-Y (2005) How phonological structures can be culturally selected for learnability, Adaptive Behavior, 13(4), pp. 269--280.

conventions like the "imitation game." In fact, the model we present here, as well as the one of de Boer, can be seen as a study of what happens when the bootstrapped vocalization repertoires generated by the agents in Oudeyer (2005b) are actually recruited for communication, which introduces feedback and repulsive forces between speech categories, and as a consequence shapes further the bootstrapped repertoires.[1]

The next section presents an overview of the artificial system. Then we show how agents form shared syllable systems efficiently and we summarize the structural properties of the produced syllable systems. Finally, we explore in detail their learnability and the implications for theories of language.

2 The Artificial System

2.1 The Imitation Game

Central to the model is the way agents interact. We use here the concept of "language game," operationally used in a number of computational models of the origins of language (Steels, 1997; Oudeyer, 2001). A game is a sort of protocol that describes the outline of a conversation, allowing agents to coordinate by knowing who should try to say what kind of things at a particular moment. Here, we use the "imitation game" developed by de Boer for his experiments on the emergence of vowel systems (de Boer, 2001).

A round of a game involves two agents, one being called the speaker, and the other the hearer. Each agent possesses a repertoire of syllables with a score associated (this is the categorical memory described below). The speaker initiates the conversation by picking up randomly one item in its repertoire and uttering it. Then the hearer tries to imitate this vocalization by producing the item in its repertoire that matches best with what it heard. The speaker then evaluates whether the imitation was good or not by checking whether the best match to this imitation in its repertoire corresponds to the item it uttered initially. It then gives a positive or negative feedback signal to the hearer. Finally, each agent updates its repertoire. If the imitation succeeded, the scores of involved items increase. Otherwise, the score of the association used by the speaker decreases and there are 2 possibilities for the hearer: Either the score of the association it used was below a certain threshold, and this item is modified by the agent who tries to find a better one, or the score was above this threshold, which means that it may not be a good idea to change this item, and a new item is created, as close as possible to the utterance of the speaker given the constraints and knowledge of the hearer at this time of its life. Regularly the repertoire is cleaned up by removing the items that have a score below a certain threshold. Initially, the repertoires of agents are empty. New items are added either by invention, which takes place regularly, or by learning from others.

2.2 The Production System

The production system that we use was designed in order to reflect the complexity of the human vocal apparatus so that we can study how some morpho-physiological constraints can affect the cultural evolution of syllable repertoires. Yet, it is not our goal to build an accurate model of the vocal tract and its associated control systems, and certainly it would not be possible since the existing conceptions of these systems in the scientific community are rather controversial and not detailed enough. Nevertheless, for other examples of vocal tract and control system models, see Bailly (1998), and Guenther (2003).

2.2.1 Vocal Tract A physical computational model of the vocal tract is used, based on an implementation of Cook's model (Cook, 1989). It consists in modeling the vocal tract together with the nasal tract as a set of tubes that act as filters, into which are sent acoustic waves produced by a model of the glottis and a noise source. There are eight control parameters for the shape of the vocal tract, used for the production of syllables. The setting of these eight parameters specifies an articulatory configuration.

2.2.2 Control System A vocalization consists in the realization of a sequence of articulatory targets. An articulatory target is defined by a specific articulatory configuration to be reached. The vocalizations that agents produce are called "syllables." Indeed, we consider the syllable from the point of view of the frame–content theory (MacNeilage, 1998) which defines it as an oscillation of the jaw (the frame) modulated by intermediary specific articulatory configurations, which represent segmental content (the content), correspond-

ing to what one may call phonemes. In our simulations, this means that agents always start a vocalization from a default rest position corresponding to a closed mouth, and always terminate the vocalization in this rest position. In between, they specify a number of articulatory targets.

There is a control system responsible for continuously driving the vocal tract shape parameters from the current articulatory configuration to the next articulatory target. This control system is simply based on a polynomial interpolation, whose smoothing properties allow us to model the phenomenon of co-articulation: The articulatory trajectory around each articulatory target, i.e., the realization of each phoneme, is modulated by the neighboring targets. This is crucial because it determines which syllables are difficult to pronounce and imitate. The constraint of jaw oscillation is modeled by a force that pulls in the direction of the position the articulators would have if the syllable was a pure frame, which means an oscillation without intermediary targets. This can be viewed as an elastic whose rest position at each time step is the pure frame configuration at this time step. The amount of deformation of the pure frame during a vocalization is used to define the articulatory cost (or energetic effort) of this vocalization. This cost is used to model the idea that easy vocalizations tend to be remembered more easily than others (Lindblom, 1992).

At the beginning of the simulation, agents are initialized with a repertoire of 10 phonemes, i.e., a set of 10 pre-defined articulatory targets, that can be thought to be the outcome of the system that we described in Oudeyer (2005b). Although the degrees of freedom that we can control here do not correspond exactly to the degrees that are used to define human phonemes, we chose values that allow them to be good analogs of vowels (V), liquids (C1) and plosives (C2), which respectively correspond to low, medium and high degree of obstruction of the air flow in the vocal tract.

2.3 The Perception System

The ear of each agent consists of a model of the cochlea, and in particular of the basilar membrane, as described in Lyon (1997). It is used to compute the evolution of the excitation of this membrane over time. Each excitation trajectory is discretized over both time and frequency: 20 frequency bins are used and a sample is extracted every 10 ms. As a measure of similarity between two perceptual trajectories, we used a technique well-known in the field of speech recognition: Dynamic time warping (Sakoe, 1982). Agents use this measure to compute which item in their memory is closest. No segmentation into "phonemes" is done in the recognition process: The recognition is done over the complete unsegmented vocalization. Agents discover what phonemes compose the syllable only after recognition of the syllable and by looking at the articulatory program associated to the matched perceptual trajectory in the exemplar.

2.4 The Artificial Brain

The brain of our agents consists of two memories of exemplars associated with a mechanism to shape and use them. The first memory consists of a set of exemplars that serve in the imitation process, and allow the agent to retrieve the motor programs which correspond to given acoustic vocalizations (this memory is therefore called the *inverse mapping memory*): The set of these exemplars represents the skills of agents for this task. This set is limited in size, which purposefully introduces a crucial generic cognitive constraint. Exemplars consist in associations between articulatory programs (an articulatory program is a sequence of articulatory targets) and corresponding perceptual trajectories. The second memory (the *categorical memory*) is in fact a subset of the inverse-mapping memory, where each exemplar possesses a score. Categorical memory is used to represent the particular syllables that count as categories in the vocalization system being collectively built by agents (corresponding exemplars are prototypes for categories). It corresponds to the memory of prototypes classically used in the imitation game (de Boer, 2001).

Initially, the inverse mapping memory is built through babbling. Agents generate random articulatory programs, composed of between one and four articulatory targets, execute them with the control module and perceive the produced vocalization. They store each trial with a probability inverse to the articulatory cost involved. The number of exemplars that can be stored in this memory is quite limited: In the experiments presented below, there are 100 exemplars whereas the total number of possible syllables is slightly above 12,000. So initially the inverse mapping memory is composed of exemplars which tend to be more numerous in zones where the cost is low than in zones where

Oudeyer, P-Y (2005) How phonological structures can be culturally selected for learnability, Adaptive Behavior, 13(4), pp. 269--280.

the cost is higher. As far as the categorical memory is concerned, it is initially empty, and will grow through learning and invention.

When an agent hears a syllable and wants to imitate it, it first looks up in its categorical memory (if it is not empty) and finds the item whose perceptual trajectory is most similar to the one it just heard. Then it executes the associated articulatory program. After the interaction is finished, it tries to improve its imitation. To do that, it finds in its inverse mapping memory the item whose perceptual trajectory matches best (it may not be the same as the categorical item). Then it tries, through babbling, a small number of articulatory variations of this item by changing, adding or deleting one articulatory target. For example, if one agent hears "ble," and if "fle" is the closest item in its memory, then it may try "vle," "fli," or even "ble" if chance decides so and if it possesses the phonemes b/l/e/f/v/i (indeed, not all possible mutations are tried, which models a time constraint: Here they typically try 10 mutations). The important point is that these mutation trials are not forgotten for the future (some of them may be useless now, but may become very useful in the future): Each of them is remembered with a probability inverse to its articulatory cost. Furthermore, this mechanism implies that the hearing of a certain kind of vocalization increases the probability to produce similar vocalizations in the future and to use them as categorical prototypes. This bears some similarities with the phonological attunement observed during the vocal babbling of developing infants (Vihman, 1996). Finally, as we have memory limitations, when new items are added to the inverse mapping memory, some others have to be pruned. The strategy chosen here is: For each new item, a randomly chosen item is also deleted (only the items that belong to the categorical memory cannot be deleted).

The evolution of the inverse mapping memory implied by this mechanism is as follows. Whereas at the beginning items are spread uniformly across "isocost" regions (they have some capacity of imitation for many kinds of vocalizations, but not very precise), at the end items are clustered in certain zones corresponding to the particular vocalization system of the society of agents (they have specific capacities for precise imitation of certain kinds of vocalizations). This is due to the fact that exemplars closest to vocalizations produced by other agents are differentiated and this leads to an increase of exemplars in their local region, and as a consequence this leads also to a decrease of the number of exemplars elsewhere, because of the limits imposed on memory. Indeed, the memory limits do not allow agents to become very good imitators in the whole space, but only in focused regions.

It is interesting to remark that what goes on in the head of each agent is very similar to what happens in genetic evolution. One can view the set of exemplars that an agent possesses as a population of individuals/genomes, each defined by the sequence of articulatory targets. The fitness function of each individual/syllable is defined by how often it leads to successful imitations when it is used, in both speaker and hearer roles. This population of individuals evolve through a generate and select process, generation being performed through a combination of completely random inventions and mutations of syllables (when one changes one articulatory goal), and selection using the scores of each syllable. The original thing here as compared to many simulations modeling either genetic or cultural evolution, is that the fitness function is not fixed but evolves with time: Indeed, the fitness of one syllable depends on the population of syllables in the heads of other agents whose fitness itself depends on this syllable. So we have a case of coupled dynamic fitness landscapes. As we will see, what happens is that those fitness landscapes synchronize at some point, they become very similar and stable. Also, the fitness of one syllable depends on the other syllables/exemplars in the memory of the agent: Indeed, if a syllable is alone in its part of the space, then few syllables of this area will be produced and other agents will have less opportunity to practice imitation of this kind of syllable, and so there is a high probability that the syllable will be pruned. The consequence of this is that group selection also happens.

3 The Formation of Shared Repertoires of Syllables

The first thing one wants to know about this system is whether populations of agents manage to develop a syllable system of reasonable size that allows them to have successful imitations. Figures 1 and 2 show an example of an experiment involving 15 agents, with a memory limit on inverse-mapping memory of 100 exemplars, and with vocalizations comprising between two and four targets included among 10 possible ones (which means that at a given moment, one agent never

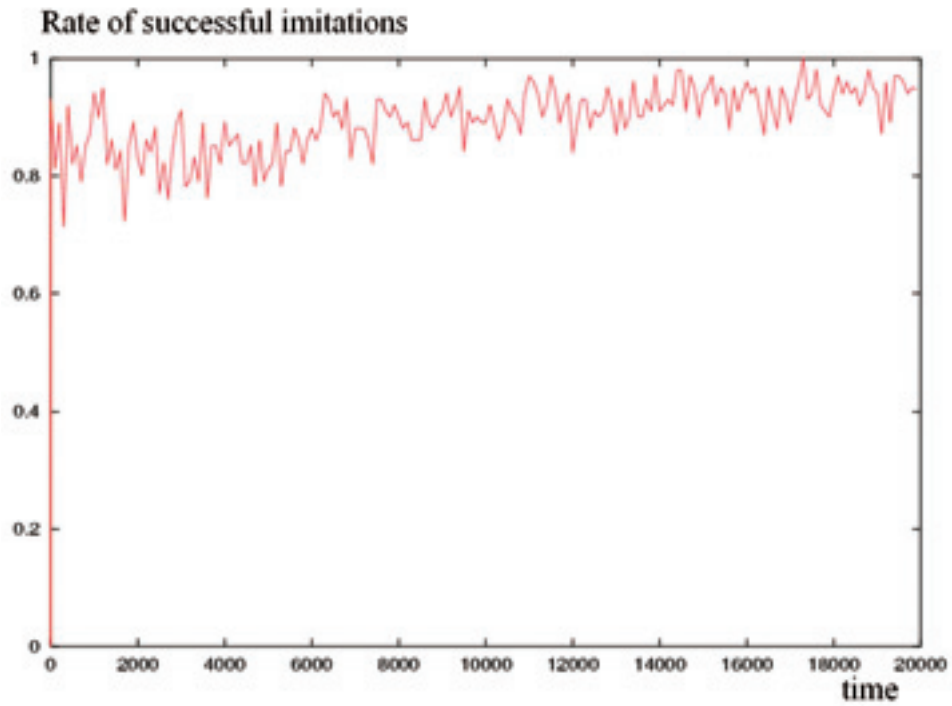

Figure 1 Example of the evolution of the rate of successful imitations in a society of 15 agents starting with empty repertoires of syllables.

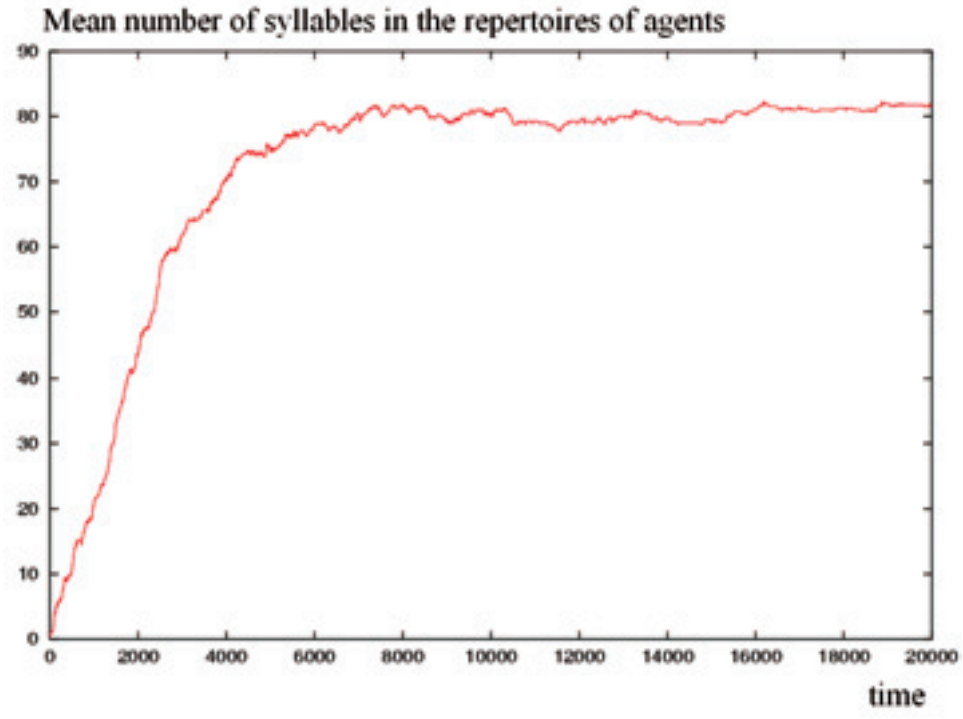

Figure 2 Corresponding evolution of the mean number of syllables in the repertoires of agents.

Oudeyer, P-Y (2005) How phonological structures can be culturally selected for learnability, Adaptive Behavior, 13(4), pp. 269--280.

knows more than about 0.8 percent of the syllable space). In Figure 1, each point represents the average imitation success in the last 100 games, and in Figure 2, each point represents the average size of categorical memory in the population (i.e. the mean number of syllables in agents' repertoires). We see that of course the success is very high right from the start: This is normal, since at the beginning agents have basically one or two syllables in their repertoire, which implies that even if an imitation is quite bad in the absolute, it will still get well matched. The challenge is actually to remain at a high success rate while increasing the size of the repertoires. The two graphs show that this is what happens. To make these results convincing, the experiments was repeated 20 times, and the average number of syllables and success was measured in the last 1,000 games (over a total of 20,000 games): 96.9% is the mean success and 79.1 is the mean number of categories/syllables. The plateau that appears in Figure 2 is directly explained by the fact that agents have a memory with limited capacity.

The fact that the success remains high as the size of repertoires increases can be explained. At the beginning, agents have very few items in their repertoires, so even if their imitations are bad in the absolute, they will be successfully recognized since recognition is done using a nearest-neighbor procedure (for example, when two agents have only one item, no confusion is possible since there is only one category). As time goes on, while their repertoires become larger, their imitation skills are also increasing: Indeed, agents explore the articulatory/acoustic mapping locally in areas where they hear the syllables of others, and the new syllable prototypes they create are hence also in these areas. The consequence is a positive feed-back loop which makes it such that agents who knew very different parts of the mapping initially tend to synchronize their knowledge and become experts in the same small area (whereas at the beginning they have skills to imitate very different kinds of syllables, but are poor when it comes to making subtle distinctions in small areas).

4 Structural Properties

It is possible to statistically study the structural properties of the self-organized syllable systems, and to compare them to human syllable systems. This was the topic of a previous paper (Oudeyer, 2001), and is not the focus of this paper. So, we only summarize the results.

The produced syllable systems have structures similar to what we observe in human languages. On the one hand, a number of universal tendencies were found, like the ranking of syllable types along their frequency (CV ≥ CVC ≥ CCV ≥ CCVC/CVCC). Also, the model predicts the preference for syllables respecting the sonority hierarchy principle, which states that within a syllable, the sonority (or degree of obstruction of the air flow in the vocal apparatus) first increases until a peak (the nucleus) and then decreases. On the other hand, the diversity observed in human languages could also be observed: Some syllable systems did not follow the trend in syllable type preference, and categorical differences exist (some syllable systems have certain syllable types not possessed by others). This constitutes a viable alternative to the mainstream view on phonological systems, optimality theory (Archangeli & Langendoen, 1997), which requires the presence of innate explicit formal constraints in the genome to account for universal tendencies (an example of constraint is the *COMPLEX constraint which states that syllables can have at most one consonant at an edge), and explains diversity by different orderings in the strengths of these formal constraints (which is basically the only thing that is learnt).

5 Learnability

The learnability of the resulting systems by new agents confronted directly with the complete vocalization system is an important question. Indeed, the learnability of language has been the subject of many experiments, theories and debates. Experiments have shown for example that language acquisition is most successful when it is begun early in life (Long, 1990), which refers to the well-known concept of the critical learning period (Lenneberg, 1967). Also, learners of a second language typically have many more difficulties than learners of a first language (Flege, 1992). Until relatively recently, these facts were interpreted in favor of the idea that humans have an innate language acquisition device (Pinker & Bloom, 1990; Piattelli-Palmarini, 1989) which partly consists in pre-giving a number of linguistically specific constraints: For example, Long (1990), argues that it is strong evidence for "maturationaly scheduled *language specific* learning abilities."

This view is also supported by a number of theoretical studies, like Gold's theorem (Gold, 1967), which basically states that in the absence of enough explicit negative evidence, one cannot learn languages belonging to the superfinite class, which includes context-free and context-sensitive languages. Nevertheless, the applicability to human languages has been challenged (Deacon, 1997).

There is an alternative view to which our model brings plausibility. It consists in explaining the fact that the learning skills of adults are lower than those of children by the fact that the brain resources needed to do so have already been recruited for other tasks or for a different language/vocalization system (Rohde & Plaut, 1999). Said another way, children learn a completely new vocalization system better than adults because their cognitive capabilities are less committed, whereas adults are already specialized. This is indeed what we can observe in the artificial system. A number of experiments were conducted in which on the one hand, some children agents (i.e. new agents) had to learn an already established syllable system, and on the other hand, adult agents had to learn the same established syllable system, which was for them a "second language" vocalization system. More precisely, in each experiment, first a society of agents was run to produce a syllable system: After 15,000 games, an agent was randomly chosen and called the teacher. This teacher was then used in the same game described above with a second agent, the learner, except that here the teacher did not update its memory (he is supposed to know that he knows the syllable system well compared to the learner). The learner was in a first series of 20 runs a child agent, and in a second series of 20 runs the learner was an agent taken from another society after 15,000 games (which models an adult who already knows another vocalization system). Typical examples of imitation success curves are in Figure 3: The upper curve is the one for child learning, and the lower curve for adult learning. Each point in the curve represents

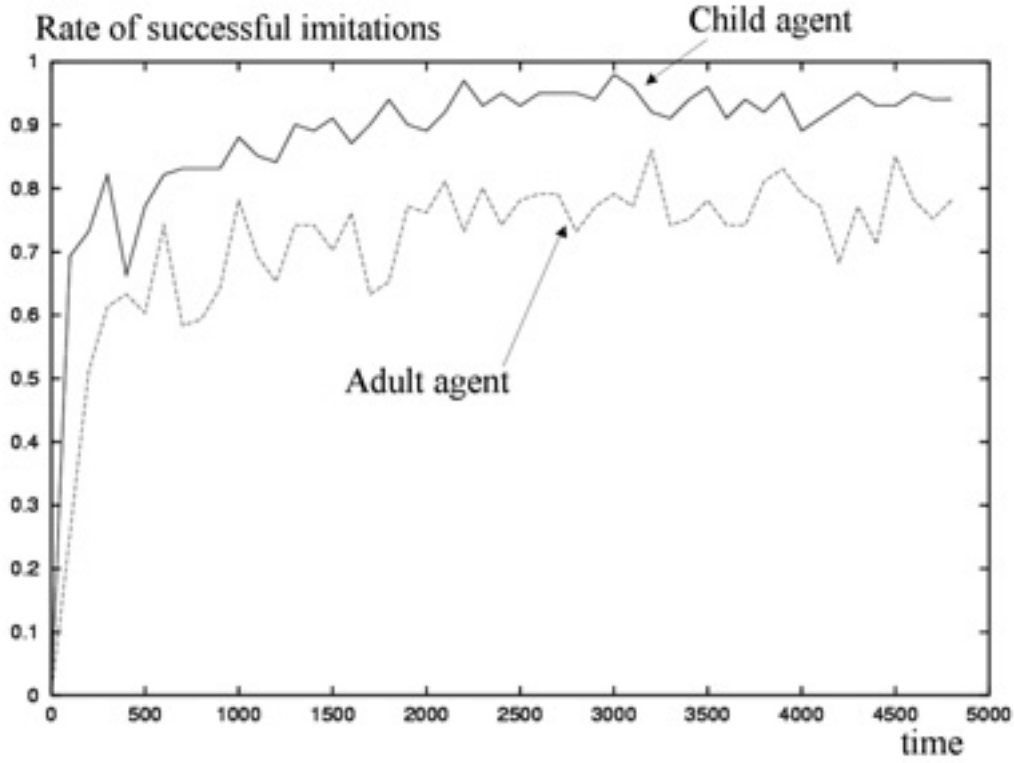

Figure 3 Evolution of the rate of successful imitations for a child agent which learns an already established syllable system (top curve), and for an adult agent already knowing another established syllable system (bottom curve). We observe that the child learns the syllable system very well, while the adult never manages to master it. This shows the "second language learning effect": adults are no longer capable of perfectly learning a new vocalization system.

Oudeyer, P-Y (2005) How phonological structures can be culturally selected for learnability, Adaptive Behavior, 13(4), pp. 269--280.

the mean success in the last 100 games at a particular time t. The mean success after 5,000 games of the 20 runs was of 97.3% for children and 80.8% for adults. This conforms well to the idea of a critical period: Adults never manage to perfectly learn another vocalization system. There is an explanation for that: Whereas children start with a high plasticity in their inverse mapping memory (because they have no categories yet and so can freely delete and create many new items) and have no strong biases towards a particular zone of the syllable space, adults, in contrast, are already committed to another vocalization system, and have more difficulty in creating new items in the appropriate zone of the syllable space because their memory resources (items in inverse mapping memory that are not prototypes of one of their previous language categories) are much lower. Of course, some of these category prototypes may be pruned, thus freeing some resources, because they are unsuccessful for the new vocalization system. But in practice it seems that a number of them allow for successful imitations with the new vocalization system, which prevents the freeing of enough resources and so the remaining confusions cannot be resolved. To conclude, we see that our model fits very well with the idea that critical periods and second language learning effect do not require a genetically programmed language-specific mechanism to find an explanation, and that the more parsimonious idea of (non-)commitment of the cognitive system can account for it.

We saw that children could actually nearly perfectly learn a fully developed syllable system. This result is not obvious since they are faced directly with the complete syllable system, as opposed to the agents who co-built it: The building was incremental and the syllable system complexified progressively, which does not mean that their job was easier since negotiation also had to take place, but it was different. An experiment was performed that on the one hand shows how non-obvious the task is and on the other hand illustrates the principle of cultural selection for learnability which we detailed in the introduction. Children agents were put in a situation of trying to learn a random syllable system: The adult/teacher was artificially built by putting in its categorical repertoire items whose articulatory programs were completely random (chosen among the complete set of combinatorially possible articulatory programs with less than four phonemes). This experiment was repeated 20 times. Figure 4 shows the curves of two experiments: The top curve is for child learning success when the target syllable system was generated by a population of agents and the bottom curve is for child learning success when the target syllable system was random. The mean success over the 20 experiments after 5,000 games is 97.3% for "natural" syllable systems and 78.2% for random syllable systems. We see that children never learn the random syllable systems well. This result is experimentally and functionally very similar to an experiment about syntax described by Christiansen (2000) in which human subjects were asked to learn small languages whose syntax was either that of an existing natural language or a random/artificial one. They found that indeed subjects were much better at learning the language where the syntax was "natural" than the language where the syntax was "artificial." Deacon (1997) also made a point about this: "If language were a random set of associations, children would likely be significantly handicapped by their highly biased guessing."

This state of affairs is in fact compatible with most of theories of language, which all basically suggest that human languages have many particular structures (that make them non-random) and that we are innately endowed with constraints that bias us towards an easier learning of these languages. Where considerable disagreement arises again about the nature and the origins of these constraints. On the one hand, the nativist approach (Pinker & Bloom, 1990) suggests that they are encoded in a Universal Grammar, genetically coded and linguistically specific, and considers language as a system mainly independent of its users (humans) who may have undergone biological evolution so as to be able to acquire and use it in an efficient way. This is not only true for syntax but also down to phonetics: This approach posits that we have an innate knowledge of what features (for example the labiality of a phoneme) and combinations of features can be used in language (Chomsky & Halle, 1968).

On the other hand, a more recent approach considers that language itself evolved and its features were selected to fit the generic and already existing learning and processing capabilities of humans (Brighton et al., 2005), and that complex linguistic structures may have emerged through a process of self-organization at multiple levels (Steels, 1997). The fact that language evolved and adapted to the primitive human brain's ecological niche, and in particular to the brains of children, explains why "children have an uncanny ability to make lucky guesses" though they do not possess innate linguistic

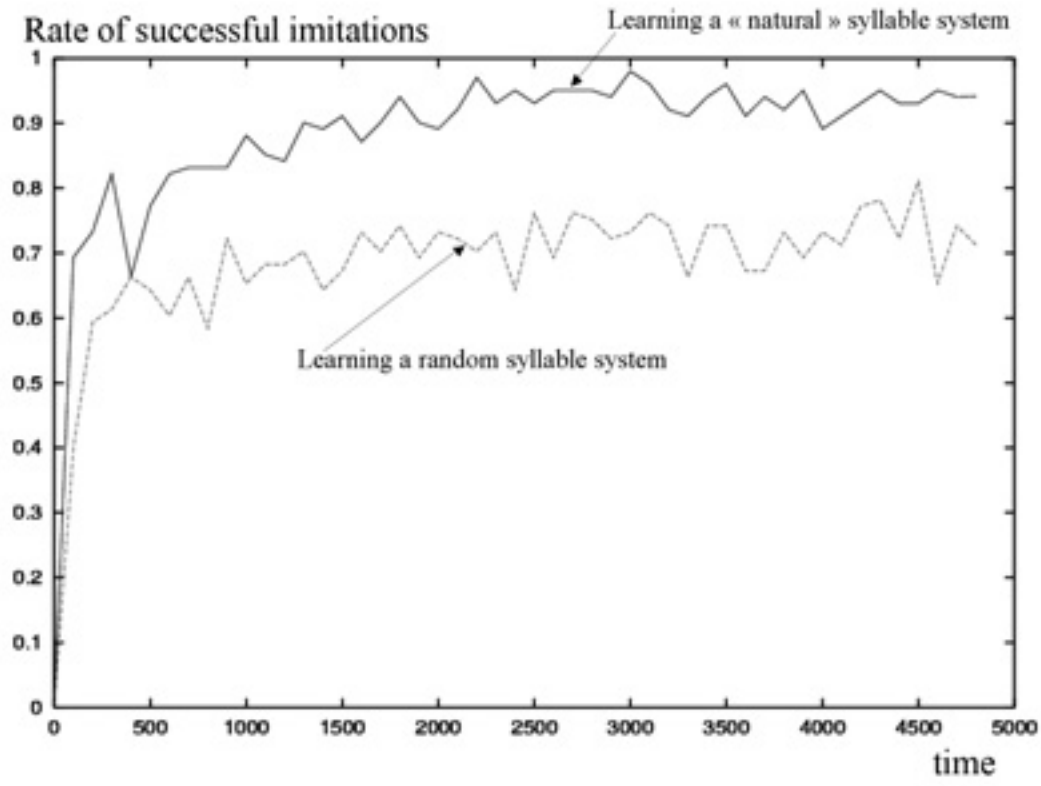

Figure 4 Evolution of the rate of successful imitations for a child agent which learns a syllable system established by a population of agents (top curve), and for a child agent which learns a syllable system established randomly by the experimenter (bottom curve). We observe the "artificial language learning effect": children can only perfectly learn the vocalization systems which evolved in a population of agents. This shows that in fact, vocalization systems were culturally selected for learnability.

knowledge (Deacon, 1997). Again the present model tends to bring more plausibility to the second approach. Indeed, it is clear here that on the one hand there are innate generic motor, perceptual and cognitive mechanisms which are not specific to speech (and could be used to learn the coordination between the vision of the hand and the muscles of the arm for example), that bias the way agents explore and acquire parts of the syllable space. On the other hand, the mechanism by which agents culturally negotiate which will be their particular syllable system makes them preferentially select systems which allow for easy imitation, hence easier learning. For instance, syllables that are very sensitive to noise will tend to be avoided/pruned since they lead to confusions and so introduce difficulties in the learning of the repertoire. Also, syllable systems will tend to be coherent both with the process of exploration by differentiation and the tendency to better remember easy syllable prototypes than difficult ones: Given a part of a syllable system, the rest may be found quite easily by focusing the exploration on small variants of items of this part, and exploration is also made maximally efficient by focusing on easy parts.

6 Conclusion

In this paper, we presented an artificial system in which agents developed shared syllable systems through cultural evolution. This system extends the one of de Boer (2001), and is complementary to the model presented in (Oudeyer, 2005a,b,c), which showed that it was possible to bootstrap complex vocalization systems without the need to pre-suppose an explicit pressure for building repertoires of distinctive sounds, and without the need to pre-suppose interactional conventions such as language games. Here, we showed how a syllable system could evolve in a population of agents once an explicit

pressure for efficient communication has been introduced, using the "imitation game" interactional framework. Moreover, this system constituted the basis for experiments which illustrated the concept of cultural selection for learnability in the context of phonology. This complements the existing simulations of this phenomenon in the field of syntax (Zuidema, 2003; Brighton et al., 2005). We showed, for example, that a child agent could easily learn a syllable system which evolved in a population of agents, but could not learn a random syllable system. Indeed, agents possess generic learning biases for which certain kinds of syllable systems are easy, and some others are difficult to learn. As a consequence, the process of cultural evolution selects the phonological systems which are easiest to learn and to transmit. These simulations develop our intuitions about the fantastic ability that children have to learn vocalization systems. In particular, this shows that an innate Language Acquisition Device which specifies explicitly the form of possible linguistic systems is not necessary to account for the performance of children. Their performance may on the contrary be due to the fact that speech evolved so as to become easily learnable. Yet, we do not exclude the possibility that biological evolution driven by the need to adapt to a linguistic environment took a role; in fact it is very probable that genes (in particular those implicated in the development of the neural system) co-evolved with language, but, as Deacon puts it: "languages have certainly done most of the adapting" (Deacon, 1997).

Note

1 Indeed, the simulations in Oudeyer (2005b,c) showed that, already, much structure can self-organize without the explicit need for building repertoires of distinctive sounds: This includes phonemic coding (the logical discreteness of continuous vocalizations), statistical regularities of vowel systems, and phonotactics.

References

Archangeli, D., & Langendoen, T. (1997). *Optimality theory, an overview.* Oxford: Blackwell.

Bailly, G. (1998). Learning to speak. Sensori-motor control of speech movements. *Speech Communication, 22*(2-3), 251–267.

Brighton, H., Kirby, S., & Smith, K. (2005). *Cultural selection for learnability: Three hypotheses underlying the view that language adapts to be learnable.* Oxford: Oxford University Press.

Cangelosi, A., & Parisi, D. (2002). *Simulating the evolution of language.* Berlin: Springer.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper Row.

Christiansen, M. (2000). Using artificial language learning to study language evolution: Exploring the emergence of word order universals in language evolution. In J. Dessalles, A. Wray, & C. Knight (Eds.), *Transitions to language* (pp. 45–48). Oxford: Oxford University Press.

Cook, P. R. (1989). Synthesis of the singing voice using a physically parameterized model of the human vocal tract. In *Proceedings of the International Computer Music Conference* (pp. 69–72), Columbus, OH.

de Boer, B. (2001). *The origins of vowel systems* (Oxford Linguistics Series). Oxford: Oxford University Press.

Deacon, T. (1997). *The symbolic species.* New York: The Penguin Press.

Flege, J. (1992). Speech learning in a second language. In C. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 565–604). Timonium, MD: York Press.

Gold, E. (1967). Language identification in the limit. *Information and Control, 10,* 447–474.

Guenther, F. (2003). Neural control of speech movements. In A. Meyer & N. Schiller (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 209–239). Berlin: de Gruyter.

Kaplan, F. (2001). *La naissance d'une langue chez les robots.* Paris: Hermes Science.

Kirby, S. (2001). Spontaneous evolution of linguistic structure—an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation, 5*(2), 102–110.

Lenneberg, E. (1967). *Biological foundations of language.* New York: Wiley.

Lindblom, B. (1992). Phonological units as adaptive emergents of lexical development. In C. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 565–604). Timonium, MD: York Press.

Long, M. (1990). Maturational constraints on language development. *Studies in Second Language Acquisition, 12,* 251–285.

Lyon, R. (1997). All pole models of auditory filtering. In E. R. Lewis et al. (Eds.), *Diversity in auditory mechanics* (pp. 205–211). Singapore: World Scientific.

MacNeilage, P. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences, 21,* 499–548.

Oudeyer, P.-Y. (2001). The origins of syllable systems: an operational model. In J. Moore & K. Stenning (Eds.). *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 744–749). Mahwah, NJ: Lawrence Erlbaum Associates.

Oudeyer, P.-Y. (2005a). *From holistic to discrete vocalizations: The blind snow-flake maker hypothesis* (pp. 68–99). Oxford: Oxford University Press.

Oudeyer, P.-Y. (2005b). The self-organization of speech sounds. *Journal of Theoretical Biology, 233*(3), 435–449.

Oudeyer, P.-Y. (2005c). The self-organisation of combinatoriality and phonotactics in vocalisation systems. *Connection Science, 17*(3), 1–17.

Piattelli-Palmarini, M. (1989). Evolution, selection and cognition: from "learning" to parameter setting in biology and in the study of language. *Cognition, 31,* 1–44.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Brain and Behavioral Sciences, 13,* 707–784.

Redford, M. A., Chen, C. C., & Miikkulainen, R. (2001). Constrained emergence of universals and variation in syllable systems. *Language and Speech, 44,* 27–56.

Rohde, D., & Plaut, D. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition, 72,* 67–109.

Sakoe, H. (1982). Dynamic programming optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 26,* 263–266.

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of Communication, 1,* 1–35.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Vihman, M. (1996). *Phonological development: The origins of language in the child.* Oxford: Blackwell.

Vogt, P. (2003). Anchoring of semiotic symbols. *Robotics and Autonomous Systems, 43*(2-3), 109–120.

Zuidema, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In S. T. Becker, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (Proceedings of NIPS'02) (pp. 51–68). Cambridge, MA: MIT Press.

About the Author

Pierre-Yves Oudeyer studied theoretical computer science at Ecole Normale Supérieure de Lyon, and obtained his PhD in Artificial Intelligence at University Paris VI, which presented a computational theory of the origins of speech sounds. He is doing research on the origins of speech, and builds societies of robots to study how linguistic codes can appear. In particular, he studies the role of self-organization in the coupling between the sensory–motor modalities within agents, and in the coupling of agents. He is also doing research in the area of developmental robotics, where he studies how a robot may autonomously set up its own tasks so that the complexity of its behavior increases in an open-ended manner. In particular, he is working on algorithms of adaptive curiosity.