

The Origins Of Syllable Systems In A Society Of Truly Autonomous Robots

Pierre-yves Oudeyer
Sony Computer Science Lab, Paris
py@csl.sony.fr

September 1, 2001

Abstract

There are many models that describe the acquisition of speech: they all rely on the pre-existence of some sort of linguistic structure in the input. Very few address the question of how this coherence and structure appeared. We propose here a solution that concerns the case of syllable systems, which are fundamental in phonology. Our model is operational and shows how a society of robotic agents, endowed with a set of non-linguistically specific motor, perceptual and cognitive constraints (some of them are obstacles whereas others are opportunities), can build collectively a coherent and structured syllable system from scratch. The structural properties of the produced sound systems are extensively studied under the light of phonetics and phonology and more broadly language theory. The model brings more plausibility in favor of theories of language that defend the idea that there needs no innate linguistic specific abilities to explain observed regularities in world languages. Results concerning the learnability of the produced sound systems by fresh/baby agents are detailed : the critical period effect and the artificial language effect can effectively be predicted by our model. The ability of children to learn sound systems is explained by the evolutionary history of these sound systems, which were precisely shaped so as to fit the ecological niche formed by the brains and bodies of these children, and not the other way around (as advocated by Chomskyan approaches to language).

1 Introduction

There are many studies about the acquisition of speech sounds, and of language in general: a lot of data is available and a lot models have been developed (Altman, 1995). Although there is a great diversity in the views they propose, one assumption underlies them all : a pre-existing structured language already exists, which is shared by a population of humans. Depending on the theoretical position, either the acquisition of a particular language or sound system consists

in adjusting a number of parameters of an innate neural structure that already knows most of the structure of languages (Chomsky and Halle, 1968), or it relies on learning techniques able to infer regularities from the data, notably through statistical learning. On the contrary, very little is known about how structures and regularities originated from a situation where there were no sound system at all (and no language in general) ?. In brief, how did speech emerge and why does it have the shape it has ? How can a shared sound system come to be used by a community of humans, and why each sound system is unique while there are general universal tendencies ? This ignorance is due partly because the general question of the origins of language has been an actively researched question only for slightly more than a decade (Hurford et al. 1998), partly because no meaningful data of sound systems of the first speaking humans exist (by nature, speech leaves no physical trace in its environment), and partly because the questions are simply very difficult. Because of this difficulty, we do not try to attack the problem of the origins of language in general, but rather to focus first on sound systems. Yet, as we will see, this sub-problem is already highly complicated.

The question of the origins is of course highly linked to the question of acquisition, which is in fact included, but adds a complete set of unsolved problems. Constraints imposed by body and cognitive characteristics or social environment constitute sometimes difficult obstacles to develop a shared repertoire of complex sounds (which might be recruited to develop a lexicon). Among these constraints for example is the decentralization of interactions among agents which have to be local in space and time, which render the only question of establishing a shared simple convention already a difficult task, which has been resolved only very recently (Steels, 1997) in the case of the collective choice of a name for a particular object in the world. Unfortunately, other constraints make this work not directly transposable to the sound systems problem : one of them is that the mapping from an articulation to a sound and then to an auditory percept is so complicated that imitating a sound, which means finding back the articulatory program that may generate the sound that has been heard, is in itself a hard job. Babies take a while before being good at imitation, but they have the advantage over primitive humans that the sounds of their environment remain stable in the course of their learning period, which is not true for primitive humans since no sound systems already existed ! Indeed, the skills that one has acquired by trying to learn the sounds invented by someone else may become obsolete, and may often require energy to be unlearned, because precisely these sounds may have been abandoned by others in the course of the negotiation, and replaced by new sounds.

There are many other such problems that face students of the origins of speech. Because the mechanisms at stake are bound to be complex and involve the interaction of many environmental, physical, neuro-cognitive and genetic entities, and because data is scarce, computational models have been increasingly used in the past 10 years in the field (Steels, 1997; Hurford et al. 1998, Kirby 1998, etc). Indeed, their nature allows on one hand to test the operational plausibility and feasibility of otherwise highly speculative theories, and on the other hand to gain new insights about how certain phenomena can be explained

by the intricate dynamics of the complex systems involved.

This article presents a computational model of the origins of syllable systems, which are thought to be a fundamental unit of the complex sound system of nowadays human languages (McNeilage, 1998). One of its aims is to prove the plausibility and feasibility of the theory that claims that sound systems originated and have properties explained by the self-organization of a number of motor, perceptual, cognitive, social and functional constraints that are not linguistically specific, and this in a cultural manner (Steels, 2000). In brief, this theory states that speech is a complex cultural adaptive system. Among the forces/constraints at stake are articulatory ease, perceptual distinctiveness, time and memory limitations, lexicon pressure, efficiency of communication, noise in the environment and conformance to the group. The word constraint is used in its most general meaning: it can be an obstacle as we suggested above, but it can also refer to opportunities as we will see. This view contrasts with the currently main stream post-structuralist conception of language, and in particular sound systems, which considers that language evolved phylogenetically and that nowadays humans have in their genome a program that develops an innate neural language acquisition device (Chomsky and Halle, 1968, Archangeli and Langendoen, 1997). This device is supposed to know already most of the formal properties of languages : for example which syntactic systems are possible or which discrete acoustic features can be used to categorize sounds (i.e. for instance the labiality or nasality of a sound). Learning plays very little in the process of learning language which consists only in setting a number of simple parameters (like the ordering of constraints in optimality theory, see Archangeli and Langendoen, 1997). Some researchers have even proposed to abandon the term “learning” for the acquisition of language (Piatelli-palmarini, 1989). Although this theory (and variants) are widely accepted, and have provided admittedly efficient (but only predictive) synchronic models of language and sound systems, they are faced with a theoretical vacuum when confronted with the questions of the origins and evolution. Indeed, these questions have been thoroughly avoided by this line of thought in the last 50 years (i.e. since the seminal work of Chomsky).

A number of computational models concerning these sound systems have already been developed, mainly for phonemes, and more precisely at the vowel level. Two of them are representative: the first one was developed by Liljencrants and Lindblom (1972) (Berrah and al., 1998 have proposed refined versions of this approach) and consisted in showing that the numerical optimisation of a number of motor and perceptual constraints defined analytically allowed to predict the most frequent vowel systems in the human languages (in particular the high occurrence of the 5 vowels system /e i a u o/). These constraints modelled basically articulatory ease and perceptual distinctiveness, which are motivated respectively by a principle of least effort and a functional need to have an efficient communication system. This work contrasted with innatist theories because it required no language specific devices to predict vowel systems that could a posteriori be described with the discrete features used by Chomskians (for example the front rounded vowels) and thought to be present in the genome. Yet, while this approach gave an idea of why vowel systems had the properties

they have, it did not give any explanation of what process could have lead to this optimisation. Indeed, it is not plausible that primitive humans may have willingly computed all possible vowel systems and took an optimal one (and still if this was the case, this does not tell us how a particular community would have agreed on one of the many solutions that exist). This shortcoming was corrected by the model of de Boer (1999, 2000) (Berrah et al. 1998, have proposed a comparable model) who places itself in a broader class of models consisting in setting up a society of agents endowed with realistic capabilities and physical constraints (whose action on the sound system is neither direct nor explicit) and have them interact in order to build culturally a efficient communication system (Steels, 1997, 1998). More specifically, in his model no explicit optimisation was performed, but rather near-optimal system were obtained only as a side effect of adaptation to the task of building a communication system. Coherence did not from a genetically pre-specified plan, but from the self-organization arising from positive feedback loops.

Much fewer existing models tackle the question of the origins of complex sounds, in particular syllables. Lindblom (1992) and then Redford (1998, 2000) have developed models resembling the Lindblom model for vowels: they consist in defining explicitly with analytical formulas a number of constraints, then running an optimisation algorithm and showing that near-optimal systems have regularities typical of the most common syllable systems in the human languages (Venneman, 1988). An example of regularity is the sonority hierarchy principle, which states that the sonority of syllables tends to first increase until their nucleus, and then decreases. The present models aims at applying the multi-agent based modelling paradigm mentioned earlier to the question of the origins and properties of syllable systems: like de Boer's model, it should not only try to explain why syllables tend to be the way they are, but also what actual process built them.

All the literature on language acquisition concerning the various modules (in particular motor and perceptual) needed by the agents of the present model are directly relevant here. Unfortunately, there exists no agreement in the scientific community on what are the right constraints at each level, so choices, had to be made since our goal is to build a complete system. This is admittedly a limit of the model, but we think that it allows a necessary exploration of how primitive humans, developing a sound systems, might have solved the problems brought by certain constraints similar in complexity and functionality to the real ones, and might have benefited from others. It also allows to understand how these constraints, which are not linguistically specific, can explain a number of phonological structures through their interactions within the complex adaptive systems formed by a decentralized population of agents who try to imitate each other.

These requirements imply the integration of many non-trivial techniques developed by various fields in artificial intelligence, speech processing and robotics. First, agents are composed of a physical model of vocal tract, which requires the simulation of aero-acoustic phenomena. Second, in order to control this vocal tract, they use a based-based module which consists in a set of simple behaviours that concurrently control each degree of freedom. Third, they are

given an artificial ear based on a model of the cochlea, with realistic electrochemical properties. Fourth, in order to compare two sounds and to cope with distortion and lack of precision in articulations, they use dynamic time warping over the trajectories of feature vectors provided by the cochlea. Fifth, their cognitive apparatus consists in exemplar based memories, which are managed with a Darwinian constructivist algorithm that has many similarities with the thesis defended by (Calvin 1987, 1996) or (Edelman 1986).

The next section describes in details these different modules. Then results about the behaviour of the system are presented concerning efficiency, structural properties and learnability properties of emergent systems. Implications for phonetics and phonology, but also more broadly for our understanding of language development and cognition in humans are discussed.

2 The model

2.1 The imitation game

Central to the model is the way agents interact. We use here the concept of game, operationally used in a number of computational models of the origins of language (Steels, 1998; Steels and Oudeyer 2000). 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.

A round of a game involves two agents, one being called the speaker, and the other the hearer. Here we just retain from their internal architecture that they possess a repertoire of items/syllables, with a score associated to each of them (this is the categorical memory described below). The speaker initiates the conversation by picking up one item in its repertoire and utters it. Then the hearer tries to imitate this sound by producing the item in its repertoire that matches best with what he heard. The speaker then evaluates whether the imitation was good or not by checking whether the best match to this imitation in his repertoire corresponds to the item he uttered initially. He then gives a feedback signal to the hearer in a non-linguistic manner. 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 he 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 to the utterance of the speaker as the agent can do given its constraints and knowledge at this time of its life. Regularly the repertoire is cleaned by removing the items that have a score too low. Initially, the repertoires of agents are empty. New items are added either by invention, which takes place regularly in response to the need of growing the repertoire, or by learning from others.

2.2 The production module

2.2.1 Vocal tract

A physical 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 8 control parameters for the shape of the vocal tract, used for the production of syllables. Finally, articulators have a certain stiffness and inertia.

2.2.2 Control system

The control system is responsible for driving the vocal tract shape parameters given an articulatory program, which is the articulatory specification of the syllable. Here 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 a segmental content (the content) corresponding to what one may call phonemes. A very important aspect of syllables is that they are not a mere sequencing of segments by juxtaposition: co-articulation takes place, which means that each segment is influenced by its neighbors. This is crucial because it determines which syllables are difficult to pronounce and imitate. We model here co-articulation in a way very similar to what is described in (Massaro 1998), where segments are targets in a number of articulatory dimensions. The difference is that we provide a biologically plausible implementation inspired from a number of neuroscientific findings (Bizzi and Mussa-Ivaldi 1991) and that uses techniques developed in the field of behavior-based robotics (Arkin 1999). This will be detailed in a forthcoming longer paper. 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. Finally, and crucially, we introduce a notion of articulatory cost, which consists in measuring on the one hand the effort necessary to achieve an articulatory program and on the other hand the difficulty of this articulatory program (how well targets are reached given all the constraints). This cost is used to model the principle of least effort explained in (Lindblom 1992) : easy articulatory programs/syllables tend to be remembered more easily than others. Agents are given initially a set of pre-defined targets that can be thought to come from an imitation game on simple sounds (which means they do not involve movements of the articulators) as described in (de Boer 1999). 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 metaphors of vowels (V), liquids (C1) and plosives (C2), which mean sonorant, less sonorant, and even less sonorant phonemes (sonority is directly related to the degree of obstruction of the air flow, which mean the more articulators are opened, the more they contribute to a high sonority of the phoneme).

2.3 The perception module

The ear of agents consists of a model of the cochlea, and in particular the basilar membrane, as described in (Lyon 1997). It provides the successive excitation of this membrane over time. Each excitation trajectory is discretized both over time and frequency: 20 frequency bins are used and a sample is extracted every 10 ms. Next the trajectory is time normalized so as to be of length 25. 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 and Chiba 1980). 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 sound. 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. This follows a view defended by a number of researchers (Seguy, Dupoux et Mehler 1995) who showed with psychological experiments that the syllable was the primary unit of recognition, and that phoneme recognition came only after.

2.4 The brain module

The knowledge management module of our agents consists of 2 memories of exemplars and a mechanism to shape and use them. A first memory (the “inverse mapping” memory) consists of a set, limited in size, of exemplars that serve in the imitation process: they represent the skills of agents for this task. Exemplars consists in associations between articulatory programs and corresponding perceptual trajectories. The second memory (the categorical memory) is in fact a subset of the inverse-mapping memory, to which is added to each exemplar a score. Categorical memory is used to represent the particular sounds that count as categories in the sound 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 1999).

Initially, the inverse mapping memory is built through babbling. Agents generate random articulatory programs, execute them with the control module and perceive the produced sound. They store each trial with a probability inverse to the articulatory cost involved ($\text{prob}=1-\text{cost}$). The number of exemplars that can be stored in this memory is typically quite limited (in the experiments presented below, there are 100 exemplars whereas the total number of possible syllables is slightly above 12000). So initially the inverse mapping memory is composed of exemplars which tends to be more numerous in zones where the cost is low than in zones where 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 sound and wants to imitate it, he first looks up in its categorical memory (if it is not empty) and find the item whose perceptual trajectory is most similar to the one he just heard. Then he executes the associated articulatory program. Now, after the interaction is finished, in any case (either it succeeded or failed), it will try to improve its imitation. To do that, it

finds in its inverse mapping memory the item (it) 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 that do not belong to the memory: each articulatory trial item is a mutated version of it, i.e. one target has been changed or added or deleted. This can be thought of the agent hearing at a point “ble”, and having in its memory the closest item being “fle”. Then it may try “vle”, “fli”, or even “ble” if the chance decides so (indeed, not all possible mutations are tried, which models a time constraints: 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 very useful in the future): each of them is remembered with a probability inverse to its articulatory cost. Of course, as we have memory limitation, when new items are added to the inverse mapping memory, some others have to be pruned. The strategy chosen here is the least biased: for each new item, a randomly chosen item is also deleted (only the items that belong to categorical memory can not be deleted).

The evolution of inverse mapping memory implied by this mechanism is as follows. Whereas at the beginning items are spread uniformly across “iso-cost” regions, which means skills are both general and imprecise (they have some capacity of imitation of many kind of sounds, but not very precise), at the end items are clustered in certain zones corresponding to the particular sound system of the society of agents, which means skills are both specialized and precise. This is due to the fact that exemplars closest to sound produced by other agents are differentiated and lead to an increase of exemplars in their local region at the cost of a decrease elsewhere.

3 Behavior of the model

3.1 Efficiency

The first thing one wants to know is simply whether populations of agents manage to develop a sound system of reasonable size and that allows them to communicate (imitations are successful). Figure 1 and 2 show an example of experiment involving 15 agents, with a memory limit on inverse-mapping memory of 100 exemplars, with vocalizations comprising between 2 and 4 targets included among 10 possible ones (which means that at a given moment, one agent never knows more than about 0.8 percent of the syllable space). In figure 1, each point represents the average success in the last 100 games, and on 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 2 graphs shows that it is the case. To make these results convincing, the experiments was repeated 20 times (doing it more is rather infeasible since each experiment basically lasts about 2 days), and the

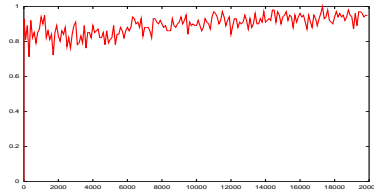


Figure 1: Example of the evolution of success in interactions for a society of agents who build a sound system from scratch

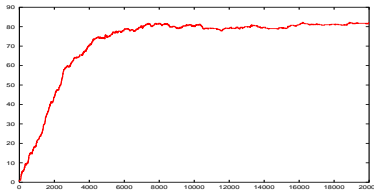


Figure 2: Corresponding evolution of mean number of items/categories in repertoires of agents along with time

average number of syllables and success was measured in the last 1000 games (over a total of 20000 games): 96.9 percent is the mean success and 79.1 is the mean number of categories/syllables.

The fact that the success remains high as the size of repertoire 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 by nearest-neighbours (for example, when 2 agents have only 1 item, no confusion is possible since their is only 1 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 other utter sounds, and the new sounds they create are hence also in these areas. The consequence is a positive feed-back loop which makes that agents who knew very different parts of the mapping initially tend to synchronize their knowledge and become expert in the same (small) area (whereas at the beginning they have skills to imitate very different kinds of sounds, but are poor when it becomes to make subtle distinctions in small areas).

This result is relevant to all theories of speech (and more generally, theories of language), innatist or not. Indeed, whereas the literature is rich of reasons explaining why having complex sound systems was an advantage for the first speaking humans is, no precise account of how it could have been built was described. For instance, even Pinker and Bloom (1990), who defend the idea that nowadays humans have a lot of linguistic knowledge already encoded in the genes, acknowledge that it certainly got there through the Baldwin effect and so was initially certainly the result of a cultural process. They give cues

Table 1: % of syllable types in produced and random systems

| | | |
|----------|----------------|----------|
| CV | CVC | CC |
| 25.3/0.2 | 20.1/1.3 | 16.1/0.5 |
| CCV | CVVC/CCVC/CVCC | other |
| 14.4/1.3 | 14.1/22.5 | 10/74.2 |

about how acquired skills and sound systems could have been transferred into the genes, but not how they got to be acquired from a situation where there was nothing !

3.2 Structural properties

Now that we have seen that a communication system was effectively built, one has to look whether the structural properties of the produced repertoires of syllables resemble human syllable systems. Indeed, human syllable systems are far from random: only very few combinations of types of phonemes occur in human languages compared to the high number of mathematically possible ones, and some occur significantly more often than others (Vennemann 1988) . For instance, all languages have CV syllables, but CVCC is rare. The difference in frequencies exist both across and within languages. A first study about the syllable types of the produced systems was achieved. Statistics about the set of all the syllables produced by 20 runs were computed (for each run, measures were done after 20000 games). Table 1 sums up the result by giving the relative frequency in use of a number of syllable types (C means “C1 or C2”). “Relative frequency in use” means that each syllable counts as the number of times it has been used by the agents in the games it played in its life. This is a better measure than simply counting the frequency of occurrence in a syllable system, because it takes into account the fact that certain syllables tend to be adopted earlier than others, which implies that they are used more times than others, and models the relative frequency effects observed within languages. The second percentage measures the proportion of the particular type of syllable in the space of all combinatorially possible ones in the experiment. This can be viewed as a measure of syllable frequencies for randomly generated repertoires.

The first observation we can do is that there is a strong difference between the relative frequencies of syllables in actual systems and in randomly generated systems. Moreover, we find that the ordering between syllables types along their frequency is very similar to the one observed in human languages (Venemann 1988), except for the presence of CC in third position (which we think is due to too low acoustic noise, unlike in the real world). These results are rather consistent with those found by Redford, and conform to what she calls the “iterative principle of syllable structure”: “simpler syllables types are expected to occur more frequently than complex ones in a systematic fashion”, where the notion of “simplicity” is constructed over the most simple syllable CV: increase in complexity comes by adding C or V iteratively at the end or beginning or

Table 2: % of syllables respecting the sonority hierarchy

| full model | jaw constraint removed | chance |
|------------|------------------------|---------|
| 70.9 per | 21.5 per | 5.3 per |

CV, and then after by replacing some C by V or the contrary.

A second important tendency of human languages is the “sonority hierarchy principle”¹ Whereas the measures in table 1 indicate that this seems to be the case in our experiment, they are too loose to conclude, especially because they blended C1 and C2. So we made measures over 20 runs about which proportion of syllables belonging to the repertoire of agents did obey the sonority hierarchy principle, using the fact that sonority of V is higher than sonority of C1, which is higher than sonority of C2 (due to the way they obstruct the air flow). Additionally, we made an experiment in which the oscillation of the jaw constraint was removed, in order to evaluate the hypothesis of Peter McNeilage that says that it is the main explanation for the sonority hierarchy principle. Table 1 sums up the results, with a column showing what is the proportion of syllable in the set of combinatorially possible syllables that respect this hierarchy.

We see that the sonority hierarchy is respected by most of the syllables of the emergent repertoires in the standard model. Yet, not all of them respect it, which is not that surprising since syllables like C1C1V do not imply an important deformation of the pure frame and so have a low cost, and do not respect the principle (there are 2 adjacent segments with the same sonority). Anyway, the actual percentage as compared to chance is much higher. When we remove the jaw constraint, we observe that the percentage of syllables respecting this hierarchy drops to around 20 percent, but is still substantially above chance. It indicates that the jaw constraint is crucial, but not the only responsible. In fact, when we remove the jaw constraint, we still start every syllable with the rest position corresponding to the closed jaw. So for instance syllables beginning with a vowel will still have a high articulatory cost. Of course for example C2C2 syllables will have a much lower cost in this case than in the case with jaw oscillation, but these syllables are very sensible to noise and do not have a high perceptual discriminability, which makes agents prune them quite often. As a result, a reasonable proportion of syllables that respect the hierarchy remain.

Until now, we have only looked at how the model produced syllable systems that reflect universal tendencies of human languages. We also have to look how well it matches with the diversity that exists across languages (Vennemann, 1988). Indeed, tendencies are just tendencies and there are cases of languages

¹ in a syllable, the sonority or loudness first increases to a peak and then may decrease again. It is very rare that for instance it first decreases and then increases or that more than 1 change in sonority direction occurs in one syllable. For instance, “ble” is preferred to “lbe”. Sonority/loudness is directly linked to the degree of obstruction of the air, and in particular to the degree of opening of the jaw.

whose syllable systems properties significantly differ from the mean (for example, in Berber, there are many syllables with long consonant sequences, and more strikingly, there are syllables that do not contain any vowel). Additionally, two languages that have for instance the same relative ratios of syllable types may implement these in very different manners. The first kind of diversity was difficult to observe in a statistically significant manner, since the relative frequencies of syllable types most often are very close to the mean above mentioned, and since not enough experiments were conducted to study rare outliers. Nevertheless, they were observed in a number of particular cases: for example, one of the obtained population had 55 percent of CVC/CCV syllables against only 20 percent of CV syllables. Some categorical differences were also observed: several populations did not have any CVVC or CVCC syllables for instance. The second kind of diversity was easier to observe in the system: you never get the same repertoires in 2 different runs of the experiment. In the 20 runs used for the experiments above, the mean number of common syllables was 20.2 (repertoires had sizes varying between 70 and 88), among which mainly 2-phonemes syllables due to their small number. Of course this result is not directly transposable to real languages since we always gave here the same set of phonemes in the beginning, whereas in reality these phonemes are not pre-given but should co-evolve with syllables, and so may lead to repertoire of syllables composed of very different phonemes ². Nonetheless, we get a good idea of how universal tendencies come from the fact that there are non-linguistically specific constraints/biases in the problem that agents are solving, whereas diversity comes from both the fact that these constraints are soft and that there exist many satisfying solutions to the problem. Operationally speaking, variety emerges because there is stochasticity locally in time and space, which makes that different societies may engage different pathways due to historical events: indeed, historicity is fundamental to the explanation of diversity. This view contrasts in different aspects with a number of innatist theories, especially optimality theory (Archangeli and Lagendoen 1997). Of course, there is a common point with optimality theory at a very general level: constraints are crucial to the explanation of language universals and diversity. Yet, a fundamental difference is the nature of constraints: in the case of optimality theory, they are linguistically specific, whereas here they are generic constraints of the motor, perceptual and cognitive apparati (we also have social constraints that are far from any concept in OT) ³. Now, the second important difference is the way these constraints are used to explain diversity: in OT, a particular syllable system corresponds to a particular ordering of constraints (some are stronger than others, which means that a low ranked constraint may be over-ridden if one has to satisfy a higher ranked constraint), which means a different constraint satisfaction problem. Conversely, in OT, one ordering of constraint implies a

² This is a limit of the model (that the model of Redford has also), but we think this limitation was necessary as a first step so that the resulting dynamics would not get too complicated to analyse.

³An example of constraint in OT is the *COMPLEX constraint which states that syllables can have at most one consonant at an edge or the NOCODA constraint which says that syllables must end with vowels.

fixed syllable system (in terms of syllables types). On the contrary, here we do not require a different set of constraints to obtain different kinds of systems, because there are many syllables systems that can be developed and allow efficient communication given only one set of constraints. Our model thus avoids a number of theoretical problems that OT is faced with: Where do the linguistic constraints come from ? If they are in the genes, how did they get there ? Why are there different orderings of constraints ? How one can pass from a set of constraints to another (which must happen since language evolves and syllable systems change) ?

4 Learnability properties

The learnability of the produced systems by fresh agents confronted directly with the complete sound is an important question. Indeed, more generally, 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 began early in life (Long, 1990), which refers to the well-known concept of critical period (Lenneberg 1967). Also, learners of a second language typically have much 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 and Bloom, 1990; Piattelli-Palmarini 1989) which partly consists in pre-giving a number of linguistically specific constraint : for example, (Long, 1990, p. 259), 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 can not learn languages belonging to the superfinite class, which includes context free and context sensitive languages (but the applicability to human languages has been challenged, see (Rohde and Plaut, 1999)).

Here we propose 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/sound system (see Rohde and Plaut, 1999 for a comparable view). Said another way, children are better to learn a completely new sound system than adults because their cognitive capabilities are less committed, whereas adults are already specialized. This is indeed what we observe in our model. To see that, a number of experiments were conducted in which on the one hand, some children agents had to learn a particular sound system, and on the other hand, adult agents had to learn a “second language” sound system. More precisely, in each experiment, first a society of agents was ran to produce a syllable system: after 15000 games, an agents was randomly chosen and called the teacher. This teacher was then used in the same game than described above, and with a second agent, the learner, except that here the teacher did not update its memory (he is supposed to know that he knows well the language as compared to the learner). The learner was each time in a first run a fresh agent (this models the child) and in a second

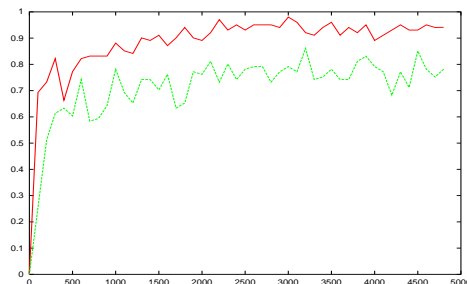


Figure 3: Evolution of success in interactions during the learning of an established sound system: top curve is when agent is a child (fresh agent) and bottom curve when it is an adult (it already knows another sound system)

run an agent taken from another society after 15000 games (which models an adult who knows already another sound system). This experiment was repeated 20 times. One example of success curve is on figure 1: the upper curve is the one for children learning success, and the lower curve for adults learning. Each point in the curve represents the mean success in last 100 games at a particular time t . The mean success after 5000 games of the 20 runs was of 97.3 percent for children against 80.8 percent for adults. This conform well to the idea of a critical period: adults never manage to learn perfectly another sound 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 bias (in fact they are biased, as we will state in next paragraph, but not as much as adults) towards a particular zone of the syllable space, adults, on the contrary, are already committed to another sound system, and have more difficulties to create new items in the appropriate zone of the syllable space because their skills resources (which are 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, and thus freeing some resources, because they are unsuccessful for the new sound system. But in practice it seems that enough of them allow successful imitations of items in the new sound system, though imprecise, so that still not enough resources can be freed to resolve the remaining confusions. To conclude this paragraph, we see that our model fits very well with the idea that critical periods/second language learning effect need not a genetically programmed language specific mechanism to find an explanation, and that the more parsimonious idea of (un-)commitment of the cognitive system can account for it.

Now, we saw that children could actually learn nearly perfectly a sound system. This result is not obvious since they are faced directly to the complete sound system, in the contrary of the agents who co-built it: the building was incremental and the sound system complexified progressively, which does not mean that their job was easier since negotiation had also to take place, but it

was different. An experiment was performed that shows on the one hand how non-obvious the task is and on the other hand has implications over a number of existing theories. Children/fresh 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 less-than-5-phonemes articulatory programs). This experiment was repeated again 20 times. Figure 2 shows the curves of 2 experiments: the top one is for child learning success when the target language was generated by a population of agents and the bottom one for child learning success when the target language was random. The mean success over the 20 experiments after 5000 games is 97.3 percent for “natural” sound systems and 78.2 percent for random sound systems. We see that children never learn reasonably well the random sound systems. This result is experimentally and functionally very similar to an experiment about syntax described in (Christiansen, 2000), in which human subjects were asked to learn small languages whose syntax was either the one 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 affair 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 biases us towards an easier learning of these languages, because they lead to the particular structure of languages. Now, where considerable disagreement comes in is again about the nature of these constraints and how they got there. On the one hand, the Chomskyan approach suggests that they are coded in a Universal Grammar genetically coded and linguistically specific, and consider 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 suggested by Pinker and Bloom, 1990, p. 712). 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 combination of features can be used in language (Chomsky and Halle 1968). One of the problems with this approach is that the apparent “idiosyncrasies of language structure are hard to explain”. On the other hand, a more recent approach considers that language itself evolved and its features were selected so as to fit to generic already existing learning and processing capabilities of humans (see for example (Christiansen 2000)), and that the coherent structures may have emerged through a process of self-organization at multiple levels (see Steels 1998). The fact that language evolved to fit to the primitive human brains ecological niche, and in particular to the brains of children, explains, as Deacon (1997) puts it, why “children have an uncanny ability to make lucky guesses” though they do not possess innate linguistic knowledge. Again the present model tends to bring more plausibility to the second approach. Indeed, it is clear here that on the one hand innate generic

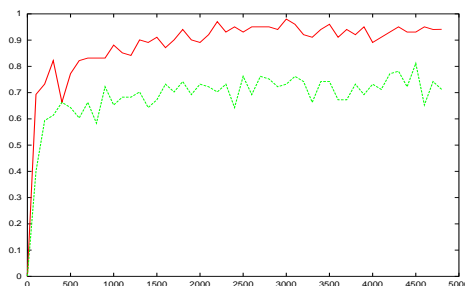


Figure 4: Evolution of success in interaction during the learning of an established sound system by a child agent: top curve is when the sound system was generated with a population of agents with all constraints, bottom curve is when the sound system is completely random

motor, perceptual and cognitive constraints bias the way one explores and acquires parts of the syllable space, and on the other hand that the mechanism by which agents culturally negotiate which will be their particular sound system makes them select preferentially systems which allow 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. Also, syllable systems will tend to be coherent both with the process of exploration by differentiation and the tendency to remember better easy items 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.

5 Conclusion

We have presented an operational model of the origins of syllable systems whose particularity is the stress on embodiment and situatedness constraints/opportunities, which implies the avoidance of many shortcuts usually taken in the literature. It illustrates in details the theory which states that language originated in a cultural self-organized manner, taking as a starting point a set of generic non-linguistically specific learning, motor and perceptual capabilities to which it adapted in order to be easily transmittable and learnable across the population and in a real and noisy environment. In addition to the demonstration of how an efficient communication system could be built with this parsimonious starting point, several specific properties that are known about human sound systems can be explained :

- universal structural tendencies like the preference for CV and CVC syllable types and the sonority hierarchy principle ;
- diversity of sound systems in spite of these universal tendencies ;

- the learnability of these sound systems by children exposed directly to their total complexity in a noisy environment and in the absence of explicit negative evidence (here only implicit negative evidence is available) ;
- the critical period effect: children learn much more efficiently a sound system than adults ;
- the artificial language learning effect (Christiansen 2000): artificial languages are much harder to learn than natural languages ;

One has to note that we do not exclude that biological evolution driven by the need to adapt to a linguistic environment played 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 done most of the adapting”.

Still, much additional experiments need to be done with this experimental setup, which should be viewed as a case-study example of how computational and robotic modelling can be used as a scientific tool for the investigation of questions concerning the cognition of humans.

6 References

Altman, (1995), *Cognitive Models of Speech Processing, Psycholinguistics and Computational Perspectives*, MIT Press.

Archangeli D., Langendoen T. (1997) *Optimality theory, an overview*, Blackwell Publishers.

Arkin, R. (1999) *Based-based Robotics*, MIT Press.

Ballard D., Hayoe M., Pook K., Rao R. (1997) Deictic codes for the embodiment of cognition, *Behavioral and Brain Sciences*, 23, 233-265.

Baldwin, M. (1896), A new factor in evolution, *The American Naturalist* 30, pp. 441-451. Reprinted in Belew and Mitchell (eds.) *Adaptive Individuals in Evolving Populations: Models and Algorithms*, SFI Studies in the Sciences of Complexity, Proc. Vol. XXVI, Addison Wesley, Reading, MA, 1996.

Banse, R. and Sherer, K. R., (1996) Acoustic Profiles in Vocal Emotion Expression, *Journal of Personality and Social Psychology*, 70(3): 614-636.

Batali, J. (1998) Computational simulations of the emergence of grammar. In Hurford, Knight and Studdert-Kennedy (eds.) *Approaches to the evolution of language: social and cognitive bases*, pp 405-426. Cambridge University Press, Cambridge.

Browman, C.P. and L., Goldstein (1992) *Articulatory Phonology: An Overview*. *Phonetica*, 49, 155-180.

Berrah A., Glotin H., Laboissire R., Bessire P., Boe L. (1998) From form to formation of phonetic structures : an evolutionary perspective, in *Proceedings of the 13th International Conference of Machine Learning, Workshop on evolutionary Computing and Machine Learning*, Venturini (ed.), Bari, Italia, pp. 23-29.

Bellman, E. (1957) *Dynamic programming*, Princeton University Press.

Berwick, R.C. (1985), *The acquisition of syntactic knowledge*, The MIT Press, Cambridge, MA.

- Bizzi E., Mussa-Ivaldi F., Giszter S. (1991) Computations underlying the execution of movement: a biological perspective, *Science*, vol. 253, pp. 287-291.
- Burmeister B., Haddadi A, Sundermeyer K., (1995) Generic, configurable, cooperation protocols for multi-agent systems, in Castelfranchi and Muller (eds.) *From Reaction to Cognition*, vol. 957, Lecture Notes in AI, 157-171, Berlin, Springer Verlag.
- Brooks R., Steels L. (1994) *The artificial life roots to artificial intelligence*, Laurence Erlbaum.
- de Boer, B. (1999) Investigating the Emergence of Speech Sounds. In: Dean, T. (ed.) *Proceedings of IJCAI 99*. Morgan Kaufman, San Francisco. pp. 364-369.
- de Boer, B. (2000) *Self-organization in vowel systems*, PhD Thesis.
- Boersma P. (1998) *Functional Phonology*, Phd Thesis.
- P. R. Cook, "Synthesis of the Singing Voice Using a Physically Parameterized Model of the Human Vocal Tract," *Proc. of the International Computer Music Conference*, pp. 69-72, Columbus, OH, 1989.
- P. R. Cook, "Identification of Control Parameters in an Articulatory Vocal Tract Model, With Applications to the Synthesis of Singing," *Electrical Engineering PhD Dissertation*, Stanford Centre for Computer Research in Music and Acoustics, Stanford University, 1991.
- Chomsky, N. and M. Halle (1968) *The Sound Pattern of English*. Harper Row, New York.
- Calvin, W. (1987), *The Brain as a Darwin Machine*, *Nature*, 330:33-34.
- Calvin, W. (1996) *The Cerebral Code*, MIT Press.
- Calvin, W. (1997) The 6 essentials ? Minimal requirements for the Darwinian bootstrapping of quality, *Journal of Memetics - Evolutionary Models of Information Transmission*, 1.
- Chomsky, N., (1957), *Syntactic Structures*, Mouton, The Hague.
- Clark A., Thornton C. (1997) Trading spaces: computation, representation and the limits of uniform learning, *Behavioral and Brain Sciences*, 20, 57-90.
- Christiansen, M., Ellefson M. (2000), Linguistic Adaptation without Linguistic Constraints: The Role of Sequential Learning, in *Language Evolution*, Dessalles, Wray, Knight (eds.), *Transitions to language*, Oxford, Oxford University Press.
- Deacon T. (1997) *The Symbolic Species*, Norton.
- Edelman, G., (1987), *Neural Darwinism, the theory of neuronal group selection*, MIT Press.
- Elman, J. (1993), Learning and development in neural networks: the importance of starting small, *Cognition* 48, 71-99.
- Flege J., (1992), Speech learning in a second language, In Ferguson, Menn, Stoel-Gammon (eds.) *Phonological Development: Models, Research, Implications*, York Press, Timonium, MD, pp. 565-604.
- Gold, E. (1967), Language identification in the limit. *Information and Control* 10, 447-474.
- Goldinger A. (1998) Echoes of echoes ? An episodic theory of lexical access, *Psychological Review*, 105, 251-279.
- Goldowsky, B., Newport E. (1993), Modelling the effects of processing limitations on the acquisition of morphology: the Less is More Hypothesis, in Clark (ed.) *The*

- Proceeding of the 24th Annual Child Language Research Forum. Centre for the Study of Language and Information, Stanford, CA, pp. 124-138.
- Greaves M., Holmback H., Bradshaw, J. (1999) What is a conversation policy ? Autonomous Agents'99 Special Workshop on Conversation Policies.
- Hardcastle, W.J. and N. Hewlett (eds.) (1999) Coarticulation. Theory, Data and Techniques. Cambridge University Press, Cambridge.
- Hurford, J., Studdert-Kennedy M., Knight C. (1998), Approaches to the evolution of language, Cambridge, Cambridge University Press.
- Johnson K. (1997) Speech perception without speaker normalization: An exemplar model. in K. Johnson, J.W. Mullenix (eds.) Talker Variability in Speech Processing, 145-165, Academic Press.
- Kirby, S. (1998), Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners, in Hurford, J., Studdert-Kennedy M., Knight C. (eds.), Approaches to the evolution of language, Cambridge, Cambridge University Press.
- Koning J-L, Oudeyer P-Y, (2000) Modelling robot-soccer strategies using conversation policies, Proceedings of ASAMA'2000, Zurich, Springer Verlag.
- Nowak, M., Krakauer D., (1999), The evolution of language, Proceedings of the National Academy of Science, USA, 96, 8028.
- Ladefoged, P. and I. Maddison (1996) The Sounds of the World's Languages. Blackwell Publishers, Oxford.
- Lenneberg, E. (1967) Biological foundations of language, New-york: Wiley.
- Lofqvist, A. (1990) Speech as Audible Gestures, in Hardcastle and Marchall (eds.) Speech production and speech modelling, pp. 289-322, Netherlands Kluwer.
- Long M. (1990) Maturational Constraints on Language Development, Studies in Second Language Acquisition 12, 251-285.
- Lindblom, B. (1992) Phonological Units as Adaptive Emergents of Lexical Development, in Ferguson, Menn, Stoel-Gammon (eds.) Phonological Development: Models, Research, Implications, York Press, Timonium, MD, pp. 565-604.
- Lindblom, B., P. MacNeilage, and M. Studdert-Kennedy (1984) Self-organizing processes and the explanation of phonological universals. In: Butterworth, G., B. Comrie and O. Dahl (eds.) (1984) Explanations for Language Universals. Walter de Gruyter, Berlin. pp. 181-203.
- Lindblom, B., and I Maddieson (1988) Phonetic Universals in Consonant Systems. In: Hyman, L. and C.Li (eds.) Language, Speech and Mind. Routledge, London, pp. 62-79.
- Liljencrants, L., Lindblom B., (1972), Numerical simulations of vowel quality systems: the role of perceptual contrast, Language, 48, pp. 839-862.
- Lyon, R. (1997), All pole models of auditory filtering, in Lewis et al. (eds.) Diversity in auditory mechanics, World Scientific Publishing, Singapore.
- McGeer, T. (1990) Passive dynamic walking, International Journal of Robotics Research, Vol. 9, No., 2, pp. 62-82.
- Mitchell, T. (1997), Machine Learning, Boston: MacGraw-Hill.
- Merzenich, M., Jenkins W. (1995), Cortical Plasticity, learning and learning disfunction. In Julesz, Kovacs (eds.), Maturational Windows and Adult Cortical Plasticity, Addison-Wesley, Reading, MA, pp. 247-272.

- Morgan J., Travis L. (1989), Limits on negative information in language input, *Journal of Child Language* 16, 531-522.
- Massaro, D. (1998) *Perceiving talking faces*, MIT Press.
- MacNeilage, P.F. (1998) The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499-548.
- Mrayati M., Carre R., Guerin B., (1988), Distinctive Regions and modes: a new theory of speech production, *Speech Communication*, 7:257-286.
- Newport E. (1990), Maturational Constraints on language learning, *Cognitive Science* 14, 11-28.
- Pinker, S., Bloom P., (1990), Natural Language and Natural Selection, *The Brain and Behavioral Sciences*, 13, pp. 707-784.
- Plaut, D. and C. Kello (1999) The Emergence of Phonology from the Interplay of Speech Comprehension and Production: A distributed Connectionist Approach. In: MacWhinney, B. (ed.) *The Emergence of Language*. Lawrence Erlbaum, Mahwah, NJ.
- Pfeifer R., Scheier (1999) *Understanding Intelligence*, MIT Press.
- 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.
- Redford, M.A., C. Chen, and R. Miikkulainen (1998) Modelling the Emergence of Syllable Systems. In: *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Erlbaum Ass. Hillsdale.
- Redford, M. A. (1999) *An Articulatory Basis for the Syllable*. Ph.d. thesis. The University of Texas, Austin.
- Redford, M. A. et al. (2000) *Constrained Emergence Of Universals and Variation in Syllable Systems, Language and Speech*.
- 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., Dynamic programming optimisation for spoken word recognition, *IEEE Transaction Acoustic., Speech, Signal Processing*, vol. 26, pp. 263-266.
- Savin, H., Bever T. (1970) The non-perceptual reality of phonemes, *Journal of Verbal Learning and Verbal Behaviour*, 9, 295-302.
- Segui, J., Dupoux E., Mehler J. (1995) The role of the syllable in speech segmentation, phoneme identification, and lexical access, in Altman, (ed.), *Cognitive Models of Speech Processing, Psycholinguistics and Computational Perspectives*, MIT Press.
- Steels, L. (1997) The synthetic modelling of language origins. *Evolution of Communication*, 1(1):1-35.
- Steels, (1998), Synthesizing the origins of language and meaning using co-evolution, self-organization and level formation, in Hurford, Studdert-Kennedy, Knight (eds.), *Cambridge University Press*, pp. 384-404.
- Steels, L. (2000) *Language as a Complex Adaptive System*, *Proceedings of Parallel Problem Solving in Nature*, Paris, *Lecture Notes in Computer Science*, Springer Verlag.
- Steels L, Kaplan F. (2000) *Talking Heads*.
- Steels, L. and R. Brooks (1995) *The Artificial Life Route to Artificial Intelligence. Building Embodied Situated Agents*. Lawrence Erlbaum, New Haven.
- Oudeyer P-y. (2001) *Coupled Neural Maps for the Origins of Vowel Systems*, *proceedings of the International Conference on Artificial Neural Networks, ICANN 2001*, Vienna, Austria.

Oudeyer P-y. (2001) The origins of syllable systems : an operational model, proceedings of the International Conference on Cognitive Science, COGSCI 2001, Edinburgh, Scotland.

Oudeyer (2000) The cultural evolution of complex sound systems without sophisticated cognitive abilities, Technical reports, Sony CSL.

Oudeyer (2001) The Expression and Recognition of emotion in speech : features and algorithms, submitted to the International Journal on Human Computer Studies, special issue on affective computing.

Oudeyer P-Y., Koning J-L, (2001) An introduction to POS: a Protocol Operational Semantics, International Journal of Cooperative Information System, Special Issue on Information Agents: Theory and Applications (2).

Perkell J. (1969), Physiology of speech production: Results and implications of a quantitative cineradiography study, The MIT press.

Thelen, E. and Smith, L. (1994) A dynamical system approach to the development of cognition and action, MIT Press.

Vennemann, T. (1988), Preference Laws for Syllable Structure, Berlin: Mouton de Gruyter.

Wittgenstein, L. (1967), Philosophical Investigations, The Blackwell Publishers.