The self-organisation of combinatoriality and phonotactics in vocalisation systems

Pierre-Yves Oudeyer

Sony CSL Paris

www.csl.sony.fr/∼py

April 6, 2005

Abstract

This paper shows how a society of agents can self-organise a shared vocalisation system which is discrete, combinatorial, and has a form of primitive phonotactics, starting from holistic inarticulate vocalisations. The originality of the system is that: 1) it does not include any explicit pressure for communication; 2) agents do not possess capabilities of coordinated interactions, in particular they do not play language games; 3) agents possess no specific linguistic capacities; 4) initially there exist no convention that agent can use. As a consequence, the system shows how a primitive speech code may bootstrap in the absence of a communication system between agents, i.e. before the appearance of language.

1 The complexity of human vocalisations

Human vocalisations have a complex organisation. They are characterized by a number of properties which need to be explained:

Discreteness and combinatoriality: speech sounds are phonemically coded as opposed to holistically coded. This implies two aspects: 1) in each language, the continuum of possible vocalisations is broken into discrete units (this is discreteness) 2) these units are systematically re-used to build higher level vocalisation structures like syllables (this is combinatoriality).

For example, in articulatory phonology (Browman and Goldstein, 1986), a vocalisation is viewed as multiple tracks in which gestures are performed in parallel (the set of tracks is called the gestural score). A gesture harnesses several articulators (e.g. the jaw, the tongue) to produce a constriction somewhere in the mouth. The constriction is defined by the place of obstruction of the air as well as the manner. While for example, given a sub-set of organs, the space of possible places of constrictions is a continuum (for example the vowel continua from low to high, executed by the tongue body) each language uses only a few places to perform gestures. This is what we call discreteness. Furthermore, gestures and their combinations, that may be called “phonemes”, are systematically re-used in the gestural scores who specify the syllables of each language. This is what we call combinatoriality. Some researchers call the combination of discreteness and combinatoriality “phonemic coding”.

Phonotactics and patterns: The way phonemes are combined is also very particular: 1) only certain phoneme sequences are allowed to form a syllable in each language, the set of which defines the phonotactics of the language (for example, “spink” is a possible syllable in English, but “npink” and “ptink” are not possible; in Tashliyt Berber, “tgzmt” and “tkSmt” are allowed, but impossible in French); 2) the set of allowed phoneme combinations is organised into patterns. This organisation into patterns means that for example, one can summarize the allowed phoneme sequences of Japanese syllables by the patterns “CV/CVN/VN”, where “CV” for example defines syllables composed of two slots, and in the first slot only the phonemes belonging to a group that we call ”consonants” are allowed, while in the second slot, only the phonemes belonging to the group that we call ”vowels” are allowed (and N stands for “nasals”).

Universal tendencies: re-occurring units of vocalisation systems are characterized by universal tendencies. For example, our vocal tract makes it possible to produce hundreds of different vowels. Yet, each particular vowel system uses most often only 3, 4, 5 or 6 vowels, and extremely rarely more than 12 (Schwartz et al., 1997a). Moreover, there are vowels that appear much more often than others. For example, most languages contain the vowels [a], [i] and [u] (87 percent of languages) while some other vowels are very rare, like [y], [oe] and [ui] (5 percent of languages). Also, there are structural regularities: for example, if a language contains a front unrounded vowel of a certain height, for example the /e/ in “pet”, it will also usually contain the back rounded vowel of the same height, which would be here the /o/ in “pot”. There are also regularities concerning the allowed sequences of phonemes. For example, all languages allow “CV” syllables, but many disallow clusters of consonants at the beginning of syllables.

Sharing: the speakers of a particular language use the same phonemes and they categorize speech sounds in the same manner. Yet, they do not necessarily pronounce each of them exactly the same way. They also share the same phonotactics.

Diversity: At the same time, each language categorizes speech sounds in its own way, and sometimes does it very differently from other languages. For example, Japanese speakers categorize the "l" of "lead" and the "r" or "read" as identical. Different languages may also have very different phonotactics.

Where does this organisation come from? There are two complementary kinds of answers that must be given (Oudeyer, 2003). The first kind is a functional answer that makes a hypothesis about the function of systems of speech sounds, and then shows that systems having the organisation that we described are efficient for achieving this function. This has for example been proposed by (Lindblom, 1992) who showed that discreteness and statistical regularities can be predicted by searching for the most efficient vocalisation systems in terms of compromise between perceptual distinctiveness and articulatory cost. This kind of answer is necessary, but not sufficient: it does not say how evolution (genetic or cultural) might have found this optimal structure. In particular, naive Darwinian search with random mutations (i.e. plain natural selection) might not be sufficient to explain the formation of this kind of complex structures: the search space is just too large (Ball, 2001). This is why there needs a second kind of answer stating how evolution might have found these structures.

In particular, this amounts to show how self-organisation might have constrained the search space and helped natural selection. This can be done by showing that a much simpler system can spontaneously self-organise into the more complex structure that we want to explain.

Self-organisation is a phenomenon complicated to understand. The computer happens to be the most suited tool for its exploration and its understanding (Steels, 1997). It is now an essential tool in the domain of human sciences and in particular for the study of the origins of language (Cangelosi and Parisi, 2002). One of the objectives of this paper is to illustrate how it can help to develop our intuitions about the role of self-organisation in the origins of language, and speech in particular.

Examples of works using this methodology have already been developed: for example (Browman and Goldstein, 2000), (de Boer, 2001), and (Oudeyer, 2001) concerning speech, and (Steels, 1997), (Kirby, 2001), (Kaplan, 2001) or (Cangelosi, 2003) concerning lexicons and syntax. As far as speech is concerned, (Browman and Goldstein, 2000) showed how the continuum of gestures could be discretized, (de Boer, 2001) showed how a society of agents could develop a shared vowel systems, and (Oudeyer, 2001), building upon the work of de Boer, showed how a society of agents could develop a shared syllable system with basic phonotactic rules. Works like (Steels, 1997; Kirby, 2001; Kaplan, 2001; de Boer, 2001; Oudeyer, 2001) provide an explanation of how a convention like the speech code can be established and propagated in a society of contemporary human speakers. They show how self-organisation helps in the establishment of society-level conventions only with local cultural interactions between agents. But they share a number of strong assumptions as far as the capabilities of agents are concerned. Indeed, the interactions between their agents follow the rules of a game which is a complex set of structured conventions. This game is called the “imitation game” in the case of (de Boer, 2001; Oudeyer, 2001). It includes for example the ability to play changing roles, to understand when one is being imitated or given a feed-back or to understand the meaning of a feed-back signal. They also share the assumption that agents are provided with the motivation to communicate and form a large repertoire of distinctive vocalisations (there are repulsive forces between the items of their repertoires). These assumptions are interesting and already permit to show a number of crucial results. But they imply that these models deal rather with the cultural evolution of languages than with the origins of language. Indeed, if one wants to understand the origins of language and speech sounds in particular, one needs to understand how the capabilities of the agents that these models assume could have appeared, which is not obvious since they are evolutionarily complex (Oudeyer, 2003).

A way to attack this question of the origins of language (speech in particular) is to show how speech codes with the above mentioned properties could be formed without such complex assumptions. The work described in (Browman and Goldstein, 2000) was a step in this direction, showing how agents who attuned the distributions of their vocalisations to each other could come to a shared discretisation of the articulatory continuum. Yet, it did study static vocalisations (these were points in an abstract one-dimensional space) and involved only two agents which self-organised a repertoire of two different vocalisations. Furthermore, the discretisation of the articulatory continuum required the presence of non-linearities in the function which mapped articulatory configurations to perceptions.

(Oudeyer, 2005) presented another system with evolutionarily simple assumptions, based on the coupling of generic neural devices which were innately randomly wired and implanted in the head of artificial agents. He showed how this system could self-organise so that the agents develop a shared vocalisation system with discreteness and statistical regularities, starting from holistic inarticulate vocalisations. The originality of the system was that: 1) it did not include any explicit pressure for communication (for example, there was no pressure to keep sounds distinctive from each other as opposed to (de Boer, 2001; Oudeyer, 2001)); 2) agents did not possess capabilities of coordinated interactions, in particular they do not play language games (as opposed to (Kaplan, 2001; Steels, 1997; de Boer, 2001; Oudeyer, 2001)); 3) agents possessed no specific linguistic capacities; 4) initially there exists no convention that agent can use (as opposed to (Kaplan, 2001; Kirby, 2001) where agents already share a system of strings or literals that they can pass to each other with no ambiguity); 5) there was no need for non-linearities in the function which maps articulatory configurations to perceptions in order to account for the discretisation of the articulatory continuum (as opposed to (Browman and Goldstein, 2000)).

This system addressed the questions of discreteness, universal tendencies of phoneme repertoires, sharing and diversity. In particular, it predicted the major statistical tendencies characterizing the vowel systems of human languages. However, it did address the question of combinatoriality only superficially, and did not address at all the questions related to phonotactics and phonological patterns. The goal of this paper is to present an extension of this system which gives an account of the systematic re-use of speech sounds in the building of complex vocalisations, and of the formation of cultural rules and patterns of sound combination. The extension is based on the addition of a map of neurons with temporal receptive fields. These are initially randomly pre-wired, and control the sequential programming of vocalisations. They evolve with local adaptive synaptic dynamics.

2 The system

We are going to make a summary of the architecture presented in details in (Oudeyer, 2005), before presenting the extension. The system is composed of agents which are themselves composed of an artificial brain connected to an artificial vocal tract and an artificial ear. Agents can produce and hear vocalisations. As described in (Oudeyer, 2005), one can model each component from the most abstract to the most realistic manner. In this paper, our goal is to explore the principles of the formation of phonotactics and of phonological patterns, rather than to build a realistic predictive model. Thus, we will use the most abstract version of the components presented in (Oudeyer, 2005). In particular, this means that agents produce two-dimensional vocalisations (one articulatory dimension and one temporal dimension). We use only one space to represent vocalisations:

the perceptual space is bypassed and only the motor space is used. So, we pre-suppose that agents can translate a vocalisation from the perceptual space to the motor space, which is acceptable since in (Oudeyer, 2005) we showed how this mapping could be learnt by the agents. The articulatory dimension that we use is also abstract, but one could imagine that it represents the place or the manner of constriction for example. Finally, the agents are put in a virtual space in which they wander randomly, and at random times they generate vocalisations which are heard by themselves as well as the closest agent.

The brain of the agent is organised into two neural maps: 1) one "spatial" neural map coding for static articulatory configurations; 2) one "temporal" neural map coding for the sequences of activations of the neurons in the static neural map (this constitutes the extension of the system presented in (Oudeyer, 2005)).

2.1 The spatial neural map

The spatial neural map contains neural units N_i which have broadly tuned gaussian receptive fields. We denote $v_{i,t}$ the centre of the gaussian related to N_i, which we call its "preferred vector" since it corresponds to the stimulus which activates maximally the neural unit. If we note $G_{i,t}$ the tuning function of N_i at time t, s one input vector, $v_{i,t}$ the preferred vector of N_i at time t, then:

$$G_{i,t}(s) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(v_{i,t}-s)^2/\sigma^2}$$

The parameter σ determines the width of the gaussian, and so if it is large the neurons are broadly tuned (a value of 0.05, as used below, means that a neuron responds substantially to 10 percent of the input space).

All the spatial neural units have initially a random preferred vector, following a uniform distribution. Each neural unit codes for an articulatory configuration, defined by the value of its preferred vector. If the neural unit is activated by the agent and a GO signal is sent to the neural map, then there is a low-level control system which drives the articulators continuously from the current configuration to the configuration coded by the activated neuron[1]. A vocalisation is thus here a continuous trajectory in the articulatory space, produced by the successive activation of some neural units in the spatial neural map, combined with a GO signal. As we will see later on, this activation is controlled internally by the temporal neurons.

As we explained earlier, we use only one space to represent vocalisations. Thus, when an agent produces a vocalisation, defined by its trajectory in the articulatory space, the agent that can perceive this vocalisation has direct access to the trajectory in the articulatory space. The perception of one vocalisation produces changes in the spatial neural map. The continuous trajectory is segmented in small samples corresponding to the cochlea time resolution, and each sample serves as an input stimulus to the spatial neural map. The receptive fields of neural units adapt to these inputs by changing their preferred vector (the width of the gaussian does not evolve). For each input, the activation of each N_i is computed, and

[1] There is always only one spatial neuron activated at a time when an agent *produces* a vocalisation, as we will explain later on. When a vocalisation is *perceived* by the agent, all spatial neurons are activated, but no GO signal is used in that case to trigger a response to the perceived vocalisation.

their receptive field updated so that if the same stimulus comes again next time, it will respond a little bit more (this is weighted by their current activation). Basically, adaptation is an increase in sensitivity to stimuli in the environment. The formula is:

$$G_{i,t+1}(s) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(v_{i,t+1}-s)^2/\sigma^2}$$

where $G_{i,t+1}$ is the tuning function of N_i at time $t+1$ after the update due to the perception of s_t at time t, and $v_{i,t+1}$ the updated preferred vector of N_i:

$$v_{i,t+1} = v_{i,t} + 0.001.G_{i,t}(s_t) \cdot (s_t - v_{i,t})$$

From a geometrical point of view, the preferred vector of each neural unit is shifted towards the input vector, and the shift is higher for unit which respond a lot than for unit which do not respond very much[2].

2.2 The temporal neural map

In (Oudeyer, 2005), the production of vocalisations was realized by activating randomly neurons in the spatial map. There was no possibility to encode the order in which the neurons were activated, and as a consequence agents ended up producing vocalisations in which all phoneme

[2] The neural network that we use is technically similar to Self-Organising Feature Maps (Kohonen, 1982). In our case, the input space is of the same dimensionality than the output space, so we do not use it to make dimensionality reduction. Feature maps are normally used to extract some regularities in high dimensional input data. Here, there is no regularity in the input data initially. Input data is generated by other neural networks of the same kind. Regularities are rather created through self-organisation as explained in the "dynamics" section.

combinations were allowed (but of course only the phonemes that appeared as a result of the self-organisation of the neural map were used). On the contrary, we will use here a temporal neural map which can encode the order of activations of spatial neurons, and is also used to activate the spatial neurons.

Each temporal neuron is connected to several spatial neurons. A temporal neuron can be activated by the spatial neurons through these connections. The tuning function of temporal neurons has a temporal dimension: their activation depends not only on the amplitude of the activation of the spatial neurons to which they are connected, but depends also on the order in which they are activated, which itself depends on the particular vocalisation which is being perceived. The mathematical formula to compute the activation of the temporal neuron i is:

$$GT_i = \sum_{t=0}^{T} \sum_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{\|t-T_j\|^2/\sigma^2} \cdot \frac{1}{\sqrt{2\pi}\sigma} . e^{\|G_{j,t}\|^2/\sigma^2}$$

with T denoting the duration of the perceived vocalisation, N the number of spatial neurons to which it is connected (which is here 2, and each temporal neuron is initially connected to 2 randomly chosen spatial neurons), T_j a parameter which determines when the temporal neuron i is sensitive to the activation of the spatial neuron j, and $G_{j,t}$ the activation of the spatial neuron j at t. Here, the T_j values are such that the temporal neuron that they characterize is maximally activated for a sequence of spatial neuron activation in which two neurons are never maximally activated at the same time and for which the maximal activation is always separated by a fixed time interval. In brief, this means that rhythm is not taken into account in this simulation: we just consider order. Mathematically,

$$T_1 = 0, T_2 = \tau, t_3 = 2 \cdot \tau, ..., T_N = (N-1) \cdot \tau$$

where τ is a time constant.

As stated in the first paragraph, the temporal neurons are also used to activate the spatial neurons. The internal activation of one temporal neuron, coupled with a GO signal, provokes the successive activation of the spatial neurons to which it is connected, in the order specified by the T_j parameters. This implies that the temporal pattern is regular, and only one neuron is activated at the same time. In this paper, each temporal neuron will be connected to only two spatial neurons, which means that a temporal neuron will code for a sequence of two articulatory targets ($N = 2$). This will allow us to represent easily the temporal neural map, but this is not crucial for the results. When an agent decides to produce a vocalisation, which it does at random times, it activates one temporal neuron chosen randomly and sends a GO signal.

Initially, a high number of temporal neurons are created (500), and are connected randomly to the spatial map with random values of their internal parameters. Using many neurons means that basically all possible sequences of activations of spatial neurons are encoded in the initial temporal neural map. The plasticity of the temporal neurons is different from the plasticity of spatial neurons[3]. The parameters of temporal

[3] Yet, some recent experiments which we do not describe in this paper because they were not conducted with the same systematicity, indicate that it is possible to use for both neural maps the same neural dynamics and still obtain results similar to those we present. In these experiments, the common neural dynamics was the same as the one we use here for the

neurons stay fixed during the simulations, but the neurons can die. As a consequence, what changes in the temporal neural map is the number of surviving neurons. The neuronal death mechanism is inspired from apoptosis (Ameisen, 2000), and fits with the theory of neural epigenesis developed by (Changeux and Danchin, 1976). The theory basically proposes that neural epigenesis consists of an initial massive generation of random neurons and connections, which are afterwards pruned and selected according to the level of neurotrophins they receive. Neurotrophins are provided to the neurons which are often activated, and prevent them from automatic suicide (Ghosh, 1996). We apply this principle of generation and pruning to our temporal neurons, and depending on their mean activity level. The mean activity of a temporal neuron j is computed with the formula:

$$MA_{j,t} = \frac{MA_{j,t-1} \cdot (window - 1) + GT_{j,t}}{window}$$

where $window$ has the initial value 50 (the value of the window size influences the speed of convergence, but the system is rather robust in terms of end result if we change it). The initial value $MA_{j,0}$ is equal to $2 \cdot vitalThreshold$. The $vitalThreshold$ constant defines the level of activity below which the neuron is pruned. This threshold remains the same for all neurons in the map. The value of this threshold is chosen so that there is not enough potential activity for all the neurons to stay alive: stability arises at the map level only after a certain amount of neurons have been pruned.

temporal neural map.

2.3 The coupling of perception and production

The crucial point of this architecture is that the same neural units are used both to perceive and to produce vocalisations, both in the spatial and in the temporal neural map. As a consequence, the distribution of targets which are used for production is the same than the distribution of receptive fields in the spatial neural map, which themselves adapt to inputs in the environment. This implies for example that if an agent hears certain sounds more often than others, he will tend to produce them also more often than others. The same phenomenon applies also to the order of the articulatory targets used in the vocalisations. If an agent hears certain combinations often, then this will increase the mean level of activation of the corresponding temporal neurons, which in turn increases their chance of survival and so increases the probability that they will be used to produce the same articulatory targets combinations. These coupling create positive feed-back loops which are the basis of the self-organisation that we will now describe.

One has to note that this is not realized through explicit imitation, defined as the repetition of a sound that has just been perceived, or of a sound that has been perceived before and has been stored explicitly in memory[4]. This is rather a side effect of an increase of the selectivity of

[4] Yet, the existence of a neural structure which allows the mapping between articulation and perception, which might correspond to the so-called "mirror neurons" (Rizzolatti et al., 1996), might still be the result of a phylogenetic evolution that happened under a selective pressure for imitation capabilities. We just say that these structures, which are only a part of a complete imitation machinery, are not used here for imitation, and their existence has the side effect of participating in the formation of a shared discrete combinatorial speech code.

neurons, and of the competition for neurotrophins between the temporal neurons, which are very generic local low-level neural mechanisms. Additionally, agents do not play any language game in the sense used in the literature (Steels, 1997). In fact, they have no capacity for coordinated protocol-based social interactions. They are just in a world in which they wander around and sometimes produce sounds and adapt to the sounds they hear around them.

3 The dynamic formation of phonotactics and patterns of combinations

In these simulations, we use a population of 10 agents. As initially the preferred vectors of the spatial neurons are random, and as there is a massive number of random temporal neurons, agents produce vocalisations which are holistic and inarticulate: the continuum of possible articulatory targets is used, and nearly all possible sequences of targets are produced. The initial state of both neural map in two agents is represented on figure 1: the spatial map is represented on the x-axis, which shows the preferred vectors, and is also represented on the y-axis, which shows the same information. The temporal map is represented by the small segments in the middle of the figure, which all correspond to a point (x, y) for which x corresponds to an existing preferred vector in the spatial map, and y to another existing preferred vector in the spatial map. The x coordinate of a temporal neuron corresponds to the first articulatory target of the vocalisation that it encodes, and the y coordinate corresponds to the second

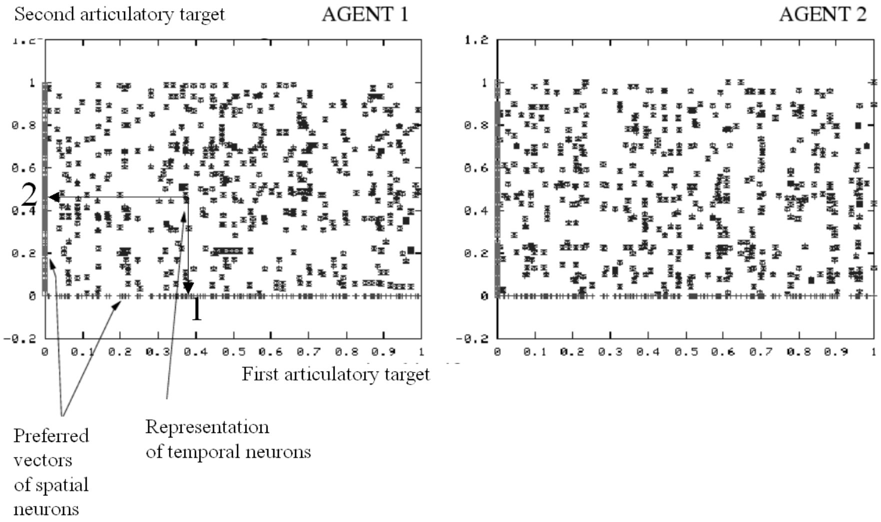

Figure 1: The neural maps of two agents at the beginning of the simulation. The neural map of one agent is represented on the left, and the neural map of the other agent is represented on the right. The spatial neurons are represented by their preferred vectors plotted on the x-axis and also plotted on the y-axis. The temporal neurons are represented by small segments (which nearly appear as points here due to their low level of neurotrophins) whose centre has its x and y corresponding to preferred vectors of the spatial neurons. The x coordinate of a temporal neuron corresponds to the first target that it encodes, and the y coordinate corresponds to the second target that it encodes.

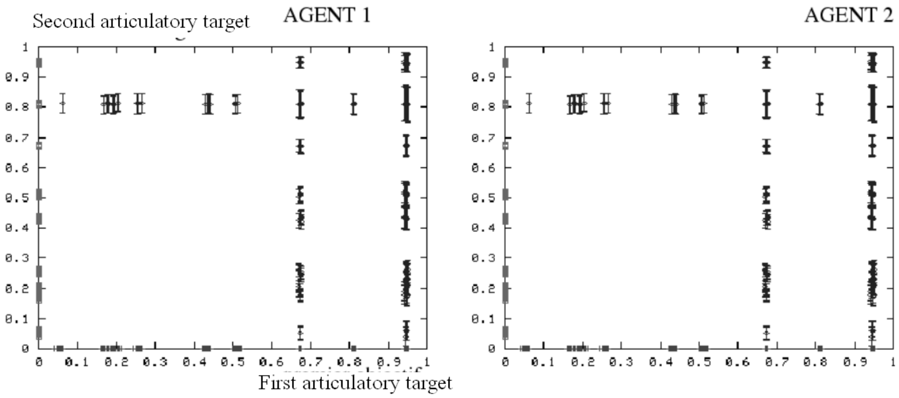

Figure 2: The neural maps of the same two agents after 1000 interactions. We observe: 1) that the preferred vectors of the spatial neural map are now clustered, which means that vocalisations are now discrete: the articulatory continuum has been broken; 2) that many temporal neurons have died and the surviving ones are organised into lines and columns: this means that phonotactic rules have appeared, that the repertoire of vocalisation can be organised into patterns, and that some phonemes get re-used systematically for building vocalisations, i.e. vocalisations are now combinatorial.

target that it encodes. The length of the segment represents the level of neurotrophins that each neuron possess.

After several hundred time steps, as we have shown and explained in details in (Oudeyer, 2005), we observe a clustering of the preferred vectors of the spatial map. Figure 2 and 3 shows an example of the neural maps after 1000 interactions in two agents (taken randomly among the 10 agents). Moreover, the clusters are the same for all the agents of the same simulation, and different for agents of different simulations. This shows that now the vocalisations that they produce are discrete: the articulatory targets that they use belong to one of several well defined clusters, and so

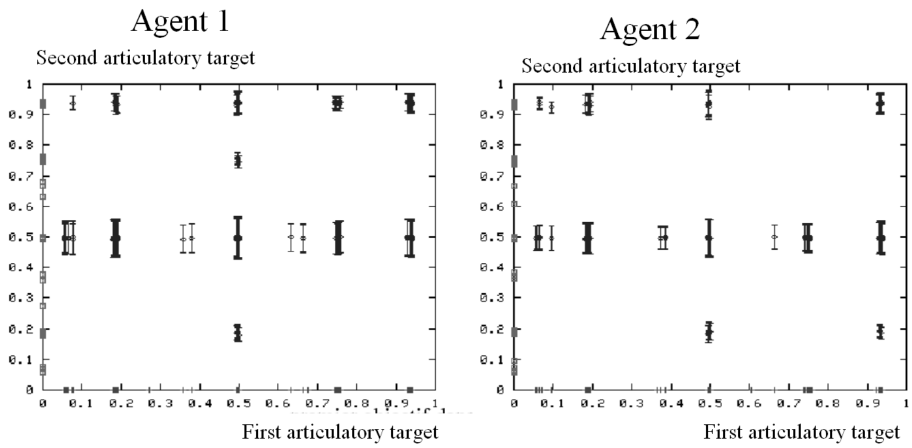

Figure 3: Another example of neural maps of two agents after 1000 interactions in another simulation.

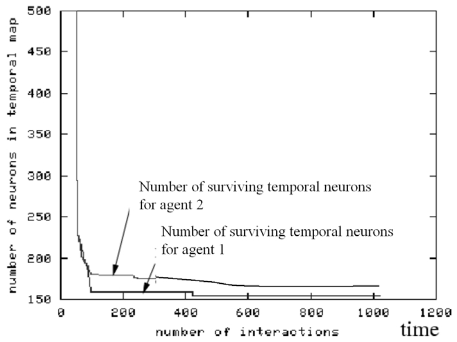

Figure 4: Evolution of the number of surviving temporal neurons corresponding to the temporal neural map of the two agents of figure 2. We observe that there is a first phase of massive pruning, followed by a stabilization which corresponds to a convergence of the system.

the continuum of possible targets has been discretized.

Moreover, if we observe the temporal map, we discover that there

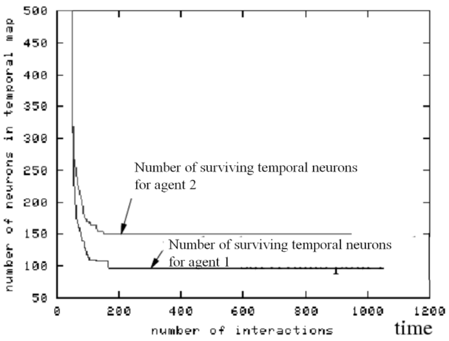

Figure 5: Another example of the evolution of the number of surviving temporal neurons, corresponding to the final neural maps of figure 3. We can observe that here the two agents do not possess exactly the same number of surviving neurons: this is due to the intrinsic stochasticity of the system. Nevertheless, as figure 3 indicates, they share the same phonotactics and the same patterns.

remains only temporal neurons coding for certain articulatory target sequences. This means that some sequences of targets belonging to the spatial clusters are not produced any more. All the agents of the same population share not only the same clusters in the spatial map, but they also share the same surviving groups of temporal neurons, as figures 2 and 3 show. This means that rules of phoneme sequencing have appeared, which are shared by all the population. In brief, this is the self-organisation of a primitive form of phonotactics. Yet, this is not all that we can observe from the temporal neural map. We also see that the surviving temporal neurons are organised into lines and columns. This means that the set of allowed phoneme sequences can be summarized by patterns. If we call the phonemes associated with the eight clusters of the spatial map on figure 2

$$p_1, p_2, ..., p_8$$

then we can summarize the repertoire of allowed sequences by:

$$(p_6, *), (p_8, *), (*, p_7)$$

where $*$ means "any phoneme in $p_1, ..., p_8$". This implies that the system of vocalisations that the agents are producing are now combinatorial: some phonemes are re-used systematically for the building of different complex vocalisations. The repertoire is thus organised into patterns. Yet, one has to remark that the types of patterns that appear are quite different from the types of patterns of real human languages, like for example the "CV/CVN/VN" organisation of syllables in Japanese. Indeed, in human languages, patterns define slots in which the set of phonemes that can appear are often disjunct: in particular, the consonants set (C) and the vowels set (V) have intrinsic properties which determine their valences and thus their privilege of occurrence in certain slots. So the complexity of the patterns that form in the simulations has not yet reached that of human languages.

The states shown on figures 2 and 3 are convergence states. Indeed, both the states of the spatial map and of the temporal neural map crystallize after a certain amount of time. In (Oudeyer, 2005), we explained in detail why the spatial map practically converged into a set of clusters for wide range of values of the parameter σ which determines the dynamics of spatial neurons.

We will now explain why there is a convergence in the dynamics of the temporal neural map, as figures 4 and 5 show (we have plotted the evolution of the number of surviving neurons within the temporal maps of two agents). As explained above, the initial level of activity ($MA_{j,0}$) of the temporal neurons is set to a constant ($2.vitalThreshold$) which is higher than the mean level of activity that will be actually computed for each neuron at the beginning of the simulation when they are still all alive. As a consequence, the mean level of activity of all neurons is going to go down at the beginning of a simulation. Because there is stochasticity in the system, due to the random choice of temporal neurons when a vocalisation is produced, and also due to the fact that all uniform distributions of preferred vectors are not exactly the same in different agents, all the $MA_{j,t}$'s will not decrease exactly in the same manner. In particular, certain temporal neurons will have their $MA_{j,t}$ go below the vital threshold ($vitalThreshold$) before the others and die (indeed,

$vitalThreshold$ is chosen so that it is higher than the mean level of activity of neurons if they are all alive). The survival of one temporal neuron in a cluster of the temporal map of one agent ag depends on the number of neurons in the corresponding cluster in other agents, whose survival depends in return on the number of neurons in the cluster of the agent ag. This creates positive feed-back loops: sometimes and by chance, a number of neurons die in the same cluster of one agent, which favours the death of similar neurons in other agents, because having less neurons in one cluster or area of the space decreases the probability to produce a vocalisation coded by the neurons of this cluster and so decreases the mean level of activity of the corresponding cluster in the other agents. Reversely, clusters composed of neurons with a high mean level of activity will favour the survival of similar clusters in other agents. This interaction between the competition and the cooperation in the clusters of temporal neurons of all agents will push a number of neurons, and a number of clusters of neurons, below the vital threshold, until there remains few enough clusters so that the neurons that compose them are activated often enough to survive and "live" together. This explains the stabilisation observed on figures 4 and 5, where we see the two phases: a first phase of initial and rapid pruning of neurons, and a second phase of stabilisation.

The "cooperation" / positive re-inforcement can happen between clusters of temporal neurons coding for the same phonemic sequence, but also between clusters of temporal neurons sharing only one articulatory target at the same location within the vocalisation. This is due to the mode of activation of temporal neurons, as detailed in the formula above. For example, let us denote p_1, p_2, p_3 and p_4 four distinct articulatory targets belonging to four distinct clusters. If the similarity of two vocalisations with the same sequence of phonemes is about 1, then the similarity between the vocalisation coded by the sequence (p_1, p_2) and the vocalisation coded by the sequence (p_1, p_3) is about 0.5, and the similarity between (p_1, p_2) and (p_3, p_4) is about 0. This means that the level of activity "provided" to the temporal neurons of a cluster *cl* thanks to two clusters of temporal neurons in other agents which share exactly one phoneme in the same location, is about the same as the level of activity provided to the neurons in *cl* thanks to the cluster in other agents which corresponds to temporal neurons sharing all the phonemes in the right location with those in *cl*. As a consequence, groups of clusters re-inforcing each other will form during the self-organisation of the temporal neurons map. These are the lines and the columns that we observed on figures 2 and 3, and this explains why we observe the formation of phonological patterns in the phonotactics developed by the agents. To summarize, the interactions between competition and cooperation among individual clusters explains the formation of shared and stable repertoires of allowed phoneme sequences, and the interaction between competition and cooperation among groups of clusters explains the formation of phonological patterns.

4 The influence of articulatory and energetic constraints on statistical preferences in phonotactics

The mechanism presented in the previous section is such that if we run a large number of simulations, there will not be any statistical preference in the localisation of clusters of spatial neurons and in the localisation of clusters of temporal neurons. We will now study how an articulatory bias can introduce preferences. As detailed in (Oudeyer, 2005), a typical articulatory bias is due to the non-linearities of the mapping between the articulatory space, the acoustic space and the perceptual space. Some small changes in the articulatory configuration of the human vocal tract can produce large changes in the acoustic and perceptual image and vice versa. If one uses an integrated architecture with one articulatory neural map and one acoustic neural map as in (Oudeyer, 2005), then even if the preferred vectors of all the neurons of both maps are initially randomly and uniformly spread across the space, their distribution quickly becomes biased by the non-linearities of the mapping (this happens if the two maps are connected so that changes in the distribution of one map are propagated to the other map, see (Oudeyer, 2005)). In (Oudeyer, 2005), we showed how the use of a realistic model of vowel perception and production could implement such a constraint and introduce statistical preferences in the repertoires of vowels formed by the societies of agents. In particular, we were able to predict the most frequent vowel systems in human languages.

Here, as we use only the articulatory representation and its associated neural map, we will model this kind of bias simply by initially generating a biased distribution of initial random preferred vectors. We chose a distribution in which there are more preferred vectors close to 1 than to 0. This is illustrated by two examples on figure 6. On the one hand, and as explained in detail in (Oudeyer, 2005), it is easy to see that this will lead to a statistical preference for clusters of spatial neurons with a preferred vector close to 1, and so for phonemes corresponding to articulatory targets close to 1. On the other hand, this bias will also influence the statistical preference of certain kinds of phoneme sequences: as there are more preferred vectors near 1 in the spatial neural map, the associated temporal neurons will be more often activated, and so their mean level of activity will be higher, which implies that they have a greater chance to survive. As a consequence, there will be a statistical preference for sequences of phonemes whose articulatory configurations of all targets are close to 1.

Using only this kind of bias is nevertheless too simplistic if one wants to grasp the principles that explain the statistical preferences for certain kinds of phonotactics over other kinds of phonotactics in human languages. Indeed, this kind of bias suggest that phonotactics preferences can be directly derived from phonemic preferences. But this is not at all the case in human languages: vowels like “a/e/i/o/u” or consonants like “t/m/n” are statistically preferred, but not all syllables sequences composed of these phonemes are statistically frequent in human languages (e.g. “ta”

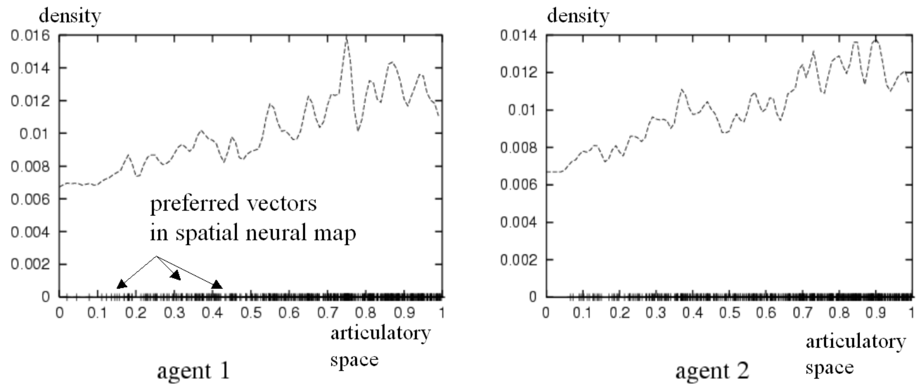

Figure 6: Example of biased initial spatial neural map: there are more preferred vectors around 1 than around 0.

or “me” are very frequent, but “tmn” or “aet” are very rare). Indeed, the statistical preferences are certainly the outcome of the interaction of several constraints.

We will illustrate this point by introducing another constraint in the system. This is an energetic constraint. In humans, each vocalisation involves the displacement of organs, which requires muscular energy: certain vocalisations are easier to pronounce from an energetical point of view than some others. Several researchers (Lindblom, 1992; Redford et al., 2001) have already proposed that this kind of energy cost was an important component in the formation of human vocalisation systems. The energy cost of one vocalisation will be modelled here as the amount of displacement of the articulator from a rest position defined as the articulatory configuration of value 0 (this is a variant of the energy cost used by (Redford et al., 2001), which measure the articulatory difference between

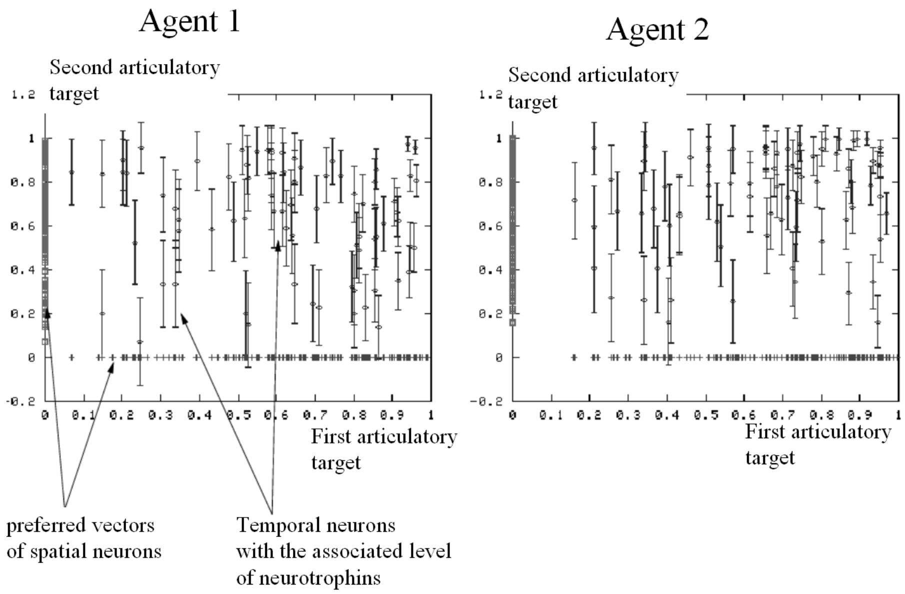

Figure 7: Example of biased initial temporal neural map: we show here the initial level of neurotrophins associated with each temporal neurons, represented by the length of the segments. We observe that temporal neurons close to $(0,0)$ have the largest initial level of neurotrophins, but that the temporal neurons close to $(1,1)$ are more numerous and so will be activated more often initially, which means that they will receive more neurotrophins than those close to $(0,0)$.

subsequent phonemes). As the speed of the articulator when it moves is here constant, there is a simple way to compute the energy associated with the vocalisation composed of the targets p_1 and p_2:

$$e(p_1, p_2) = p_1^2 + p_2^2$$

This energy will influence the survival of the temporal neurons. Indeed, we explained earlier that the survival of temporal neurons depended on the level of neurotrophins that they received. A neuron could receive neurotrophin in proportion to its level of activation. The stress associated with the spending of energy can in reverse prevent the reception of neurotrophins (Ghosh, 1996). In particular, temporal neurons coding for vocalisations with targets close to 0 will be favoured by this constraint as compared to the temporal neurons coding for vocalisations with targets close to 1. We will denote $Nt_{i,t}$ the level of neurotrophins received by the temporal neuron N_i at time t. Then we can compute:

$$Nt_{i,t} = MA_{i,t} - c_1.e(p_{1,N_i}, p_{2,N_i})$$

where c_1 is a normalizing constant so that both the terms of activation and of energy have the same ranges, and where p_{1,N_i} and p_{2,N_i} are respectively the first and second articulatory target encoded by temporal neuron N_i. Again, there is a constant $vitalThreshold$ such that if the level of neurotrophins $Nt_{i,t}$ becomes smaller, then the temporal neuron N_i is pruned. This constant is chosen so that not all temporal neurons can survive. Here, $MA_{i,0} = 0.06$, $vitalThreshold = 0.03$, $c_1 = 15$ and there are 150 spatial neurons and 500 initial temporal neurons. Figure 7 gives two examples of initial spatial and temporal neural maps. The segments

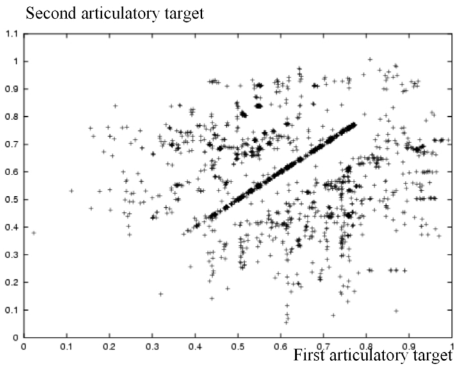

Figure 8: Distribution of surviving temporal neurons in 500 simulations.

on the representation of temporal neurons represent here the initial value of their neurotrophin level $Nt_{i,0}$. As the $MA_{i,0}$ component is the same for all temporal neurons, this gives also a representation of the energy cost associated with the vocalisations coded by the temporal neurons (the higher the segment, the lower the energy cost). We represented here only 100 temporal neurons instead of 500 for a better visibility. This figure shows that there are more temporal neurons with associated targets close to 1, but that these neurons with targets close to 1 have individually the lowest level of neurotrophins.

We are now going to run the system and observe how the combination of these two constraints can lead to the formation of phonotactic systems whose statistical properties can not be deduced from each constraint studied independently. We ran 500 simulations, made a database of all the surviving temporal neurons after convergence of the system, and plotted them on the figure 8. We observe that there is a clear statistical preference for vocalisations composed of targets located in the centre of the space, and not near 0, as the energetical constraint alone would result, or near 1, as the non-linearity articulatory constraint alone would result.

This shows how crucial it is to understand in detail all the constraints influencing the formation of repertoires of vocalisations, as well as the interaction among these constraints, if one wants to understand why for example human languages prefer CV syllables to CCVC syllables. This result is positive in the sense that it illustrates the kind of dynamics that can give rise to apparently idiosyncratic phonotactics regularities. This helps us develop our intuition of the self-organised processes which shape vocalisations systems. But this result is also negative in the sense that it shows how far we are from being able to predict human languages statistical preferences in phonotactics. Indeed, our knowledge of the physiological, energetical and representational dimensions of human speech is extremely low. There are few areas for which we have probably good models, such as the perception and production of vowels, which allow to use realistic constraints in a predictive model of the statistical regularities of vowel systems (de Boer, 2001; Oudeyer, 2005). But for example we know very few about the energetical cost of vocalisations, and the existing models of the brain representations of speech signals, which is crucial for the understanding of the articulatory/perceptual non-linearities, are still very speculative. We are not even able to make a list of all the possible constraints that might influence the process of creation of vocalisations. This also explains why instead of building a system based on very speculative models of realistic constraints, we chose to build a system with completely abstract representations and constraints, which facilitates the understanding of the dynamics.

Finally, it should be said that another constraint which would be very interesting to integrate is the functional constraint. Indeed, for reasons that we explained in the introduction, we developed a system free of functional constraint: the agents had no motivation for building a communication system with a repertoire of distinctive vocalisations. We showed (Oudeyer, 2005) that even without this motivation, and with no repulsive force, still the system self-organised a shared repertoire of vocalisations which can be categorized distinctively. Yet, if we imagine that this system actually describes a process that took place in the evolution of humans before they had language, it was certainly recruited later on in order to communicate. This means that a functional pressure came in and added new constraints, such as the perceptual distinctiveness between similar vocalisations, which typically would disfavour sequences of identical phonemes like "aaa" or "mmm". This case could be studied by coupling the system described in this paper with the imitation game invented by (de Boer, 2001) and extended to syllables in (Oudeyer, 2001).

5 Conclusion

In (Oudeyer, 2005), we presented a system showing how a society of agents could self-organise a discrete speech code shared by all speakers of the same community, and different in different communities. We also showed how it allowed to predict certain statistical regularities characterizing the repertoires of phonemes in human languages. The originality of the system was that: 1) it did not include any explicit pressure for communication; 2) agents did not possess capabilities of coordinated interactions, in particular they did not play language games like the "imitation game"; 3) agents possessed no specific linguistic capacities; 4) initially there exists no convention that agents can use; 5) there was no need for non-linearities in the function which maps articulatory configurations to perceptions in order to obtain the discretisation of the articulatory continuum.

This made the system a good tool to think and develop our intuitions about the bootstrapping of speech, and attack the problem of the origins of language as opposed to the problem of the formation of languages which has already been studied extensively in the computer modelling literature (e.g. Kaplan (2001); Kirby (2001); de Boer (2001); Oudeyer (2001); Cangelosi and Parisi (2002)). Indeed, by making evolutionarily simpler assumptions than existing models, it allows us to understand how natural selection, in an environment favouring the reproduction of individuals capable of communication, could have been guided by self-organisation to establish the first and primitive forms of conventions, such as the speech codes that our agents generate. In this paper, we presented a natural and crucial extension to our earlier work, introducing a mechanism that takes into account the order of articulatory targets both in production and in perception of vocalisations. This allowed to show that similarly, agents could self-organise combinatoriality and a primitive form of phonotactics defining shared sets of allowed phonemic sequences in a given population. Diversity was again a feature: different populations of agents developed different phonotactics systems. Moreover, the set of allowed phonemic sequences could always be organised into patterns. Yet, these patterns are quite different from the types of patterns of real human languages in which there are phonological categories like consonants and vowels, which possess disjunctive valences and privileges for the occurrence in certain syllabic slots. Searching for mechanisms which could account for the formation of such phonological categories will be the subject of future work.

We also studied theoretically how the addition of constraints such as non-linearities due to the articulatory/acoustic mapping or such as the energetic cost of vocalisations could influence the statistical preferences of populations of agents for certain kinds of phonotactics. This showed that if one wants to be able to predict the actual phonotactics preferences in the human languages, then it is crucial to take into account all the constraints as well as their interactions. Unfortunately, the speech sciences are too young and our knowledge of these constraints is today either speculative or not detailed enough. Whereas it is possible to make relatively realistic models of the production and the perception of vowels, which allows to build predictive models of human vowels systems (e.g. Lindblom (1992); Schwartz et al. (1997b); de Boer (2001); Oudeyer (2005)), existing models of the production and perception of consonants, and models of the production and perception of sequenced articulatory targets can hardly be used in a predictive model of human phonotactics because they would introduce too much ad hoc and speculative biases. Indeed, let us take the example of the “Frame-Content Theory” developed in (MacNeilage, 1998), which states that vocalisations consist in the deformation of “frame” cycles of opening and closing the jaw, and thus that vocalisations are subject to the articulatory cost of the deformation of these default cycles. This theory, even if it provides interesting insights into the understanding of speech, does not specify operationally how this cost is computed by the motor system and the neuronal networks to which it is connected: as a consequence, the potential modeller is left with the obligation to invent cost functions, and this will necessarily introduce assumptions which will have a strong impact on the result of a simulation, as we have shown in this paper. There is thus a risk that these assumptions, not founded on real observations, distort the initial qualitative theory (e.g. the "Frame-Content Theory") and destroy the potential benefits of using it in a simulation.

This is also why we preferred to stay at an abstract and theoretical level in the work that we presented in this paper, which has the advantage of allowing to understand better the biases which are programmed in, but also to understand the biases that could be introduced by for example a so-called "realistic" model of the vocal tract. Because of these considerations, we believe that the priority in the possible continuations of this work is not to introduce realistic models of the human perceptual and production apparatus for complex vocalisations, but to study the incorporation of a functional pressure for communication. Indeed, we showed here that one can already go a long way without a functional pressure for communication, but if one wants to bridge the gap with the formation and evolution of contemporary speech systems, it is a necessity to use such a pressure. Indeed, some phenomena can only be accounted with it, like the existence of large vowels systems (more than 10 vowels) which requires a mechanism of active phonemic creation and repulsive forces among the different phonemic categories. This study could be done by coupling the system which we presented in this paper with higher level systems like the ones described in (de Boer, 2001) or (Oudeyer, 2001).

6 Acknowledgements

I would like to thank the anonymous reviewers for their precious comments on this paper. This research has been partially supported by the ECAGENTS project founded by the Future and Emerging Technologies programme (IST-FET) of the European Community under EU R&D contract IST-2003-1940.

References

Ameisen, J., 2000. La Sculpture du vivant. Le suicide cellulaire ou la mort cratrice. Seuil.

Ball, P., 2001. The self-made tapestry, Pattern formation in nature. Oxford University Press.

Browman, C., Goldstein, L., 1986. Towards an articulatory phonology. Phonology Yearbook 3, 219–252.

Browman, C., Goldstein, L., 2000. Competing constraints on intergestural coordination and self-organization of phonological structures. Bulletin de la Communication Parle 5, 25–34.

Cangelosi, A., 2003. Neural network models of category learning and language. Brain and Cognition 53 (2), 106–107.

Cangelosi, A., Parisi, D., 2002. Simulating the evolution of language. Springer Verlag.

Changeux, J., Danchin, A., 1976. The selective stabilization of developing synapses: a plausible mechanism for the specification of neuronal networks. Nature 264, 705.

de Boer, B., 2001. The origins of vowel systems. Oxford Linguistics. Oxford University Press.

Ghosh, A., 1996. Cortical development: With an eye on neurotrophins. Current Biology 6, 130–133.

Kaplan, F., 2001. La naissance d' une langue chez les robots. 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.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biological Cybernetics 43 (1), 59–69.

Lindblom, B., 1992. Phonological units as adaptive emergents of lexical development. In: Ferguson, Menn, Stoel-Gammon (Eds.), Phonological Development: Models, Research, Implications. York Press, Timonnium, MD, pp. 565–604.

MacNeilage, P. F., 1998. The frame/content theory of evolution of speech production. Behavioral and Brain Sciences 21, 499–511.

Oudeyer, P.-Y., 2001. The origins of syllable systems : an operational model. In: Moore, J., Stenning, K. (Eds.), Proceedings of the 23rd Annual Conference of the Cognitive Science society, COGSCI'2001. Laurence Erlbaum Associates, pp. 744–749.

Oudeyer, P.-Y., 2003. L'auto-organisation de la parole. Ph.D. thesis, Université Paris VI.

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

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

Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L., 1996. Premotor cortex and the recognition of motor action. Cognitive Brain Research 3, 131–141.

Schwartz, J., Bo, L., Valle, N., Abry, C., 1997a. Major trends in vowel systems inventories. Journal of Phonetics 25, 255–286.

Schwartz, J., Boe, L., Valle, N., Abry, C., 1997b. The dispersion/focalization theory of vowel systems. Journal of phonetics 25, 255–286.

Steels, L., 1997. The synthetic modeling of language origins. Evolution of Communication 1.