The cultural evolution of syntactic constraints in phonology.

Luc Steels(1,2) and Pierre-Yves Oudeyer (1)
(1) Sony Computer Science Laboratory - Paris
(2) VUB Artificial Intelligence Laboratory - Brussels
steels@arti.vub.ac.be

Abstract

The paper reports on an experiment in which a group of autonomous agents self-organises through cultural evolution constraints on the combination of the individual sounds (phonemes) in their repertoires. We use a selectionist approach whereby a repertoire evolves by mutations of patterns, constrained by functional pressures from perception and production and the need to conform to the group.

Introduction

Language was commonly viewed in the 19th century, including by Charles Darwin, as a living system which evolves in a cultural fashion. This changed with the structuralist movement in linguistics that dominated research in the 20th century. Structuralism emphasises the formal description of language as an idealised system at a specific moment in time, which is largely innate. This approach has therefore not produced significant explanatory formal models on how language has emerged or how it evolves. Principles and modeling techniques from artificial life research can make a major contribution, although they need to be applied to cultural rather than genetic evolution. This paper reports on a case study in the cultural evolution of a particular nontrivial aspect of language, namely phonology.

The sound system of a natural language like English is constrained in two ways: The repertoire of individual sounds (phonemes) that speakers of a particular language are able to produce, recognise, and reproduce are a subset of all the possible sounds that the human vocal apparatus can in principle produce (Ladefoged and Maddieson,1996). For example, English does not use the vowel [y] (pronounced as in French "rue") whereas French does. Second, a language *constrains* the set of possible sound combinations. For example in English [mb] can occur at the end of a word as in "lamb" but not in the beginning, whereas in some African languages this is possible (as in Swahili "mbali" (far)). Sounds fall into classes and the classes form a combinatorial system. It follows that the emergence of the sound system of a specific language has two aspects: (1) the emergence of a repertoire of individual sounds and (2) the emergence of an additional level of syntactic complexity constraining their combination. This paper is concerned with the problem how phonological classes and combinatorial constraints may emerge and continue to evolve. It is a concrete case study on how a level of (syntactic) complexity and systematicity might self-organise from independent units through cultural evolution.

The emergence of constraints on sound combinations requires that (1) sounds become grouped into classes, and (2) that combinatorial constraints among members of these classes become conventionalised in the population. Three theories have been put forward to explain the origin and acquisition of such phonological constraints. The most widely accepted theory at the moment, which developed from structuralist research in the line of Jakobson, and Chomsky and Halle (1968), is that the categorisation of sounds is based on a set of innate distinctive features (like voiced, fricative, etc.) and that their combinatorial constraints are a subset of universal combination principles chosen by setting some parameters. This suggests a genetic origin of phonology and a maturational approach to language acquisition (see e.g. Dresher, 1992). The second theory takes an empiricist track and proposes that the sound system of a language is acquired through an inductive statistical learning process which delineates classes based on the distribution of sounds in the inputs given to the learner. Most research within a connectionist framework follows this line (see e.g. Plaut and Kello, 1999).

This paper adopts a third alternative which is based on selectionist principles. It proposes that there are mechanisms in each agent that generate in a basically random fashion possible sound systems and variations on sound systems, but that the set of possibilities is constrained by two selectionist forces: There are functional constraints coming from production and recognition, for example, the sound "wrljts" is much more difficult, if not impossible, to reproduce easily and to recognise reliably compared to a sound like "baba". Self-organisation due to a positive feedback loop between use and success acts as a secondary selectionist force to ensure that speakers of the same language share the same conventions. This selectionist hypothesis has been put forward by a number of authors (Lindblom, MacNeilage, and Studdert-Kennedy (1984), Steels (1997a)) and has been suggested for the evolution of grammatical complexity as well (Hashimoto and Ikegami (1996), Kirby (1999), Steels (1997b)).

Our research group has developed a general framework for exploring this selectionist approach, not only for speech (De Boer, 1999) but also for the origins of lexicons (Steels and Kaplan, 1999) and grammar (Steels, 1997b). The framework assumes a population of distributed autonomous agents that take turns playing a consecutive series of games. Each game exercises some aspect of language (sound production and sound recognition in the case of experiments in phonetics) and is followed by adaptation based on feedback from the outcome of the game. For investigating speech, we have been employing imitation games in which the speaker produces a random sound from his repertoire, the hearer recognises the sound and attempts to reproduce it. Feedback is based on the speaker's judgement whether the hearer's sound is indeed the one the speaker produced. Adaptation includes the adoption of a new sound, shifting of a sound (in perceptual or production space), or elimination of a sound from the repertoire. Speakers may occasionally create new sounds by adopting a new randomly chosen configuration of the articulators. So far it has not only been shown that a repertoire of sounds (albeit only vowels) can emerge from such games but also that the possible repertoires satisfy the tendencies observed universally in human vowel systems as long as the speech apparatus and the hearing system are reasonably realistic with respect to human speech (De Boer,1997). Some preliminary work has been done on syllables (Redford, et.al. 1998) but not yet through multi-agent simulations.

This paper adopts the same framework for studying the emergence of sound combinations, more specifically syllables, like "pa", "bri", "art", etc. A typical language has about 250 to 300 possible syllables which are then combined into higher order units like words. At this point we have only studied the problem formally, i.e. by assuming an abstract articulatory space, and an abstract perceptual space. This way we can study more generally how complex units may form from simple ones in a collective self-organising process. But the abstractions do not take away the major problems that need to be dealt with:

- *The inverse mapping problem.* The key difficulty in acquiring a sound repertoire is to learn how to move the articulators to reproduce a particular sound, based only on acoustic information about the sound. Because the articulatory space has many degrees of freedom which do not map directly onto the dimensions of the perceptual space this problem cannot be solved analytically, even if a good physical model would be available to the language learner. The inverse mapping problem is exacerbated by the fact that a smooth transition between positions in articulatory space may lead to non-smooth transitions in perceptual space.
- *Combinatorial explosion.* The number of possible syllables exponentially increases with the size of a repertoire (a 20 phoneme repertoire gives rise to 160,000 possible combinations of size 4 for example), so an exhaustive search for viable combinations is excluded.
- *Contextual influence.* Sounds are influenced by the context due to coarticulatory side-effects (Hardcastle and Hewlett, 1999). For example, a vowel before a nasal consonant (as in "on") is already slightly nasalised. Even the [k] in the word "cow" already shows the first signs of the lip-rounding associated with [w]. Consequently sounds in isolation are acoustically different from sounds in combination. In fact, some consonants cannot even be pronounced without context.
- *Continuous parameters* The distinctive features traditionally used in abstract phonology cannot be assumed as given. The parameters controlling the articulators are continuous. For example, the horizontal position of the tongue can go from high to low. Different languages carve up this continuum in different ways. The data from perception is continuous as well, unsegmented and uncategorised. A realistic simulation should therefore include a mechanism for mapping features onto the speech signal, rather than assuming that phonological features are given.
- *Memory limitation.* It is unrealistic to assume that agents store large sets of examples to which they can repeatedly return, as is done in many statistical learning approaches. We need case by case online learning without memory of past cases. Storing

sounds or gestures in full detail is to be avoided as well as it would require enormous memory resources.

We have found it useful to decompose the evolution of phonological complexity into three transitions. For each transition we discuss (1) what is needed in terms of cognitive architecture to enable the increased complexity, (2) what results we obtained in simulating the transition, and (3) what selectionist pressure justifies the additional complexity. The three steps are: from individual (static) sounds to complex (dynamic) sounds, from complex undifferentiated sounds to sound patterns, and from sound patterns to categorial constraints.

From static to dynamic sounds

Individual phonemes

Our starting point are earlier simulations, in particular those by De Boer (1997), which clearly demonstrate the viability of the selectionist approach for individual isolated static sounds. The simulations assume a population of agents, which can be changing if we want to research questions of language transmission or language contact, that play a consecutive series of imitation games. Each agent is capable to store a repertoire of sounds in an associative memory. A sound has two components: (1) A target in the articulatory space, for example, for the sound [u] as in "boot", the horizontal tongue position is towards the back, vertical tongue position is up towards the roof, and lips are rounded. (2) A region (with a prototypical midpoint) in acoustic space, made up by the sound's energy level within certain frequency bands, known as formants. For example, the [u] sound is typically found in a region around F1=276 Hz, F2=740 Hz and F3=2177 Hz (Vallee, 1994). Agents must be able to control their articulatory apparatus to reproduce a sound in their repertoire and they can recognise which sound was produced based on a similarity match between the signal heard and the sound's corresponding region in acoustic space.

For each imitation game, two members are selected randomly from the population. The first agent acts as speaker, the second as hearer. The speaker chooses one sound from his repertoire and puts his articulators in the prescribed position, thus generating an acoustic signal. The hearer perceives the signal in terms of his perceptual space and retrieves the sound that is closest to the signal. Then the hearer reproduces his version of the sound which is perceived and categorised again by the speaker. If the speaker agrees that this is the same sound, he gives a positive feedback otherwise a negative one.

Based on this feedback the agents update their associative memory. When the sound could be correctly recognised and reproduced, their respective scores go up. Otherwise, the hearer may either move the perceived sound closer to the one heard by hill-climbing (both in acoustic and articulatory space) or add a new one when the sound heard was too far from the ones existing so far in his repertoire. A score is kept of the use and success of sounds. Those sounds that consistently fail or do not occur frequently enough are discarded. The computer simulations carried out by de Boer (1997) have abundantly shown that a collective repertoire of individual vowels indeed self-organises through these mechanisms and that the emerging repertoires exhibit the same characteristics as those found in natural phonologies, specifically they occupy preferentially the extrema of the vowel space and then start filling up the spaces in between.

A key property of a selectionist approach is that perceptual analysis only has to be able to *differentiate* the sounds that are effectively in the repertoire because this is enough to retrieve the motor program producing the sound. This requires less fine-grained feature extraction than if the position of the articulators has to be recovered for an inverse-mapping. For example, in Japanese no firm distinction exists between [l] and [r] so that Japanese speakers can be (and are) less sensitive than English speakers for this distinction. At the same time, imitation has to be only as precise as required to distinguish the different sounds. An imitation may sound inadequate for the ears of the speakers of one language, whereas it is perfectly fine from the perspective of another language.

Complex sounds

The major limitation of the simulations of de Boer is that only individual static sounds are employed, so that the problem of co-articulatory effects and non-linearity due to sequencing of articulatory targets do not appear. However, human languages make obviously use of more complex, connected sounds as opposed to individual sounds. Hence the speaker must be able to produce an articulatory gesture (Browman and Goldstein, 1992) and thus generate a continuous sound signal in time. The signal appears as a trajectory in the acoustic space of the hearer (usually called a signature). The main reason why languages use complex dynamic sounds is because the set of individual sounds reliably producable by the human vocal apparatus is limited and so this would severely restrict the semantic potential of a language (individual speakers have vocabularies of at least 50,000 words). Many sounds (particularly the consonants such as [t]) are almost impossible to produce and reliably perceive in isolation. So the pressure to develop a broader repertoire of phonetic building blocks pushes the emergence of more complex sounds, and we believe that this has inevitably given rise to the development of a complex phonology as we show in the paper.

But let us start first from a situation where the agents use complex sounds but no phonology yet. This means that a complex dynamic sound is taken as an individual unit *in toto*. A sound in the associative memory of the agents then consists on the one hand of a complete articulatory gesture, and on the other hand of a complete signature in acoustic space. Selectionist forces coming from reproducability and perception are again at work to restrict the set of possible sound combinations (as discussed by Lindblom and Maddieson, 1988). Agents now need a more complex matching function to find the sound that is similar to the ones in their repertoire. The matching function needs to be complex because it includes a time aspect, and because due to the difficulty of articulation, speaker variance, influence of environmental noise, and other factors such as the variable speed of speech, two acoustic signatures for the same sound will never be exactly the same.

Oudeyer (1999) has performed a series of imitation game experiments from this perspective, using exactly the same population dynamics as in the de Boer experiments. A realistic synthetic articulator built by Eduardo Miranda based on the Cook synthesiser (Cook, 1989) and real signal processing was used. It was shown that imitation was not harder than with static sounds and that a repertoire of shared sounds could emerge. This may seem a paradoxical result, because there are now many more degrees of freedom (14 for articulation). But in some sense the task is easier because when the articulatory space contains many more dimensions the agents have a less hard time to find regions that can be reliably be distinguished. Hence two complex sounds which humans perceive as being very different might nevertheless be considered as successful imitations by the agents.

These simulations showed at the same time the strong limitations of this approach:

- *Memory usage*: A new pair (articulatory gesture, acoustic gesture) needs to be stored for every complex sound in the language and it needs to be stored with full articulatory or acoustic precision. This is obviously not tenable from the viewpoint of memory storage and retrieval.

- *Repertoire size*: When the sound repertoire becomes larger, the acoustic space and the articulatory space become more crowded, making a more precise recognition and more accurate reproduction necessary. The inverse-mapping problem then becomes more prominent and agents have a much harder time to collectively self-organise a shared repertoire.

- *Language acquisition*: The language learner is required to learn every sound separately, which puts a strong limit on the speed of language acquisition. It leads to difficulty in language acquisition which can be observed in the simulations because new sounds have a hard time to propagate in the rest of the population.

Oudeyer (1999) has shown that these difficulties impose strong limitations on the repertoire size that can be generated and maintained by a population. Specifically, the repertoire could never get beyond 20 complex sounds. So when lexicon and grammar pressure for a wider repertoire, the system will not be able to deliver. Each of the three dimensions above needs to be improved: Some form of drastic compression is needed to keep memory usage within bounds. Reproduction and recognition somehow need to handle much tighter regions in articulatory and acoustic space, and the spreading of new sounds must become faster. We believe that these three forces provide the positive selection pressure for the emergence of a phonological system.

From complex sounds to sound patterns

This brings us to the second transition. It consists in breaking up the elements of a complex sound into components which each have the properties of individual sounds discussed earlier: a target in articulatory space and a target in acoustic space. Such a break-up results in an enormous compression. The trajectory between the articulatory targets can be filled in by the motor system and need no longer be stored. The perceptual system gets the main targets but can ignore what happens in between. Once a complex sound is broken up in individual sounds, the same individual sound can be re-used in other complex sounds, giving an additional compression. Such an approach is also beneficial from the viewpoint of language acquisition because once an individual sound has been learned, it can be used to recognise an unknown complex sound.

Handling Patterns

To achieve this transition, the speech memory of the agents must be pattern-based, a realistic assumption born out by various types of psychological evidence (MacNeilage,1998). A pattern consists of a sequence of slots and possible fillers of each slot (figure 1). Patterns are widely used throughout the brain for various

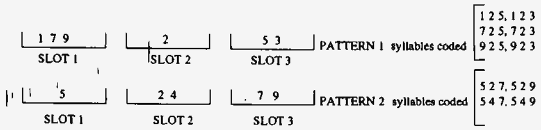


Figure 1: Two example patterns are shown together with the syllables they cover. Individual sounds are denoted by numbers.

tasks and could therefore easily have been recruited for speech. A specific realisation of a sound pattern corresponds to a syllable, such as "bla". Other syllables that could be governed by the same pattern are: "pla", "blo", "bli", "pli", etc. The possible fillers of a slot are individual sounds (phonemes) which are stored in a sound-memory similar to the one used in the imitation games discussed in the previous section. Each sound is an association between a point in articulatory space (an articulatory target) and a region with prototypical midpoint in acoustic space. Each slot-filler pair in a pattern has a score, decomposed into the use and success of the filler in the pattern. An agent's phonetic repertoire consists of a set of sounds, a set of sound patterns, and a specification what sounds can fill which slot in each pattern.

When an agent produces a complex sound (a syllable), he first selects a pattern and then selects a sound for each slot in the pattern. The sounds provide a series of consecutive articulatory targets (goals), starting from the rest state. The way the agent tries to achieve these targets is similar to other motor control tasks, like skiing down a slope with targets set on the way, and so we have used similar behavior-based techniques as those used for example in mobile robotics (Steels and Brooks (1995)) to plan and execute articulatory gestures. Each target acts as an attractor pulling the articulators involved towards it. At the same time an articulator exhibits inertia which slows down the approach to the target (for example the tongue can only move at a certain speed) and takes into account feedback towards the goal which repels the movement towards the target when there is a risk of overshooting. These different dynamical forces acting on a trajectory are captured in the following equations

$$pos(t+1) = pos(t) + f(c_1 * attract(t) +$$
$$c_2 * feedback(t) + c3 * inertia(t)$$
$$attract(t) = \sum_{i/t_i>t}(-c_4(t_i - t) + c_5)\frac{pos(t)g_i}{norm(pos(t)g_i)}$$
$$feedback(t) = pos(t)g_i - \frac{t_i - t}{TIMESTEP}pos(t-1)pos(t)$$
$$inertia(t) = pos(t-1)pos(t)$$
$$f(v) = if(norm(v) > MAXSPEED)$$
$$then\ (MAXSPEED\frac{v}{norm(v)})\ else\ v$$

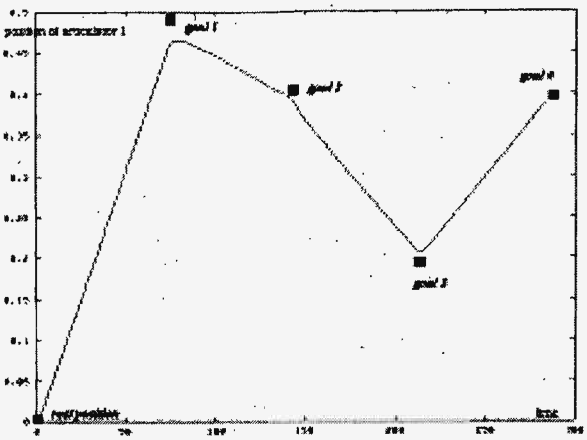


Figure 2: This figure shows the trajectory of one articulator for a pattern consisting of four slots.

These equations determine the position (pos) of an articulator at time t (in milliseconds), given $(g_0, t_0)(g_1, t_1)$... where each g_i is an articulatory target and t_i the timing of the target. $c_1 ... c_5$ and $MAXSPEED$ are parameters of the model and $norm(v)$ is the norm of vector v.

An example of a trajectory (for only one single articulatory dimension) is shown in figure 2. It has been produced using the same realistic articulatory synthesiser as for the Oudeyer experiment discussed in the previous section. There are 14 trajectories for each of the articulatory dimensions.

Due to the inertia factor there is influence of the previous target on the next and due to the fact that the next target already starts pulling, there is anticipation just as in human speech. The articulatory synthesiser introduces additional co-articulatory effects that show up in the final signal. Realising a trajectory is not a matter of simply touching each target one after the other. Often there is no time to reach a target or a target can only be approached slightly before moving on to the next one. For instance, we see that the trajectory in figure 3 does not completely touch goal1. Sometimes two targets are so far apart that a trajectory is not possible. This puts strong constraints on the set of viable sound combinations that can be used by the agents. Agents can be seen as experimentally determining which sound combinations 'work' from an articulatory point of view.

The articulatory trajectories enacted by the speaker generate a trajectory in the perceptual space of the hearer. Figure 3 shows the outcome of the trajec-

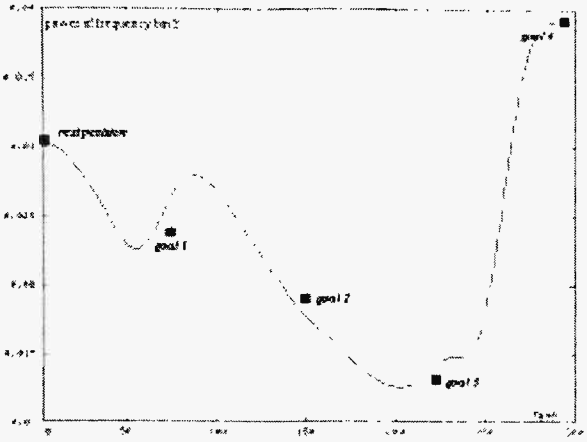


Figure 3: Trajectory of one of the acoustic dimensions produced by the articulatory trajectory partially represented in Figure 2.

tory partially described in figure 2, after considerable smoothing. A phonetic event is defined as a significant change of the first derivative in an acoustic trajectory. For each phonetic event a possible corresponding sound is retrieved by comparing its point in acoustic space to those stored with the individual phonemes in memory. Because the articulatory gesture seldom creates a perfect path, the acoustic signature can never be expected to yield a perfect match. Occasionally the match will be so bad that it is not possible to recognise the individual sounds.

The trajectory in figure 3 illustrates this quite clearly. The different acoustic goals of the individual sounds have been superimposed. Goal1 is more or less reached although a bit earlier than expected. Goal2 does not show up as a significant phonetic event, i.e. significant change in the direction of the trajectory, at all, instead there is a phantom event (due to nonlinearity) between goal1 and goal2 which does not correspond to any articulatory target and hence cannot be recognised as a sound. Goal3 can be recognised although there is another confusing phonetic event close to it. Goal4 has been reached completely. The difficulty of aligning articulatory targets with acoustic targets puts a second strong constraint on the set of possible sound combinations: The individual targets in the pattern must be recognisable despite the distortion caused by combining them. The particular sequence of goals in figure 2 and 3 is an example of a non-viable pattern, because of the non-linear phenomena between goal1 and goal2.

Once individual phonemes have been recognised, the hearer needs to find the syllable in his pattern memory that matches the series of phonemes. Assuming that the hearer has found a pattern (or possibly more than one), he then reconstructs the articulatory gesture by setting the relevant articulatory targets and re-synthesising the complex sound. When the pattern did not exist yet in the hearer's repertoire, the hearer signals ignorance but nevertheless attempts to reproduce the syllable based on recognition of the individual sounds in the perceived syllable. The speaker then in turn attempts to recognise the syllable produced by the hearer based on his own repertoire and gives positive or negative feedback depending on whether a matching pattern could be found or not.

The agents adapt their memory based on the outcome of the imitation game. The scores of the slot-fillers that were used go up in the case of full success. In case of imitation success but cultural failure (the hearer did not possess the syllable although he could imitate it based on stringing together individual phonemes), the speaker decreases the scores of the slot-fillers in the pattern that was used and the hearer tries to incorporate the pattern in his own repertoire. There are two conditions: (1) the existing pattern must match sufficiently close (which means that there is only a difference in one slot), and (2) the new slot-filler must be compatible with the other slot-fillers already in the pattern. If this is not the case a totally new pattern is constructed if there is memory available. Finally, if imitation was not successful at all, only the speaker updates his memory by lowering the scores of the slot-fillers involved.

Occasionally the speaker generates a new pattern by a random combination of sounds from his sound repertoire, or mutates a pattern by trying out another slot-filler than those used so far. As only a subset of all possible combinations is viable (either from the viewpoint of perception or production) there is no guarantee that a randomly selected sound may be fittable in a pattern and so 'natural selection' from perception or production weeds out these patterns. Moreover because speaker and hearer locally exchange information whether a pattern is shared, the population pressure acts as secondary selectionist force on what repertoire will form. The population can be seen as performing a collective search for a shared set of patterns through a process that is similar to spin-glass style relaxation. Agents have a strong limitation on the set of possible patterns they can store in memory, so that they are forced to prune patterns to make way for new ones. The pruning criteria are based on how many different slot-fillers a pattern has and on the score of the slot-fillers.

As in a genetic system or in the immune system, mutation rates need to be regulated. Mutation gives rise to innovation and therefore to expansion of the repertoire. But when mutations are too rapid, the syllables cannot spread in the population and so imitation success starts to drop. The mutation rate has therefore been coupled to success. An agent stops mutating patterns when the success rate is not sufficiently high (below 85 %), so as to give a chance to absorb new syllables or have new syllables be absorbed by the rest of the population. Of course, if the agents have to lower their mutation rate the increase in the repertoire goes less fast. So this is a way to measure the efficiency of the language acquisition process.

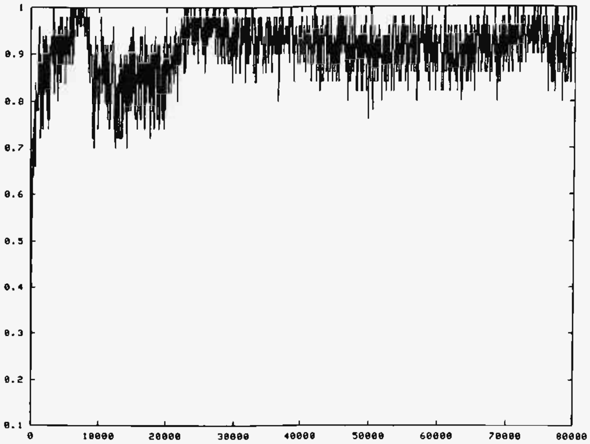


Figure 4: Each point represents the percentage of success for 50 games with 30 agents

Simulation Results

The mechanisms skeched above have been integrated in a computer simulation. An example of the results obtained with this simulation is shown in the following figures. These results are for a population of 30 agents which does not change. Agents start with a repertoire of 10 individual sounds which are derived using isolated sound imitation games, as in the de Boer experiment, and their objective is to construct a shared repertoire of sound patterns with these sounds. The parameters (e.g. maximum speed of articulators or timing between two goals) are set such that about 50 percent of syllables of length between three and five are not viable from an articulatory or perceptual point of view. Figure 4 shows the success percentage over time every 50 games and figure 5 the increase in the size of the repertoire.

There are four phases. In the first phase, success rapidly rises to almost 100 percent success. This is a phase where the patterns have basically one slot-filler so there is little ambiguity how they have to fit together. The situation is similar to learning individual syllables. However, in phase 2, the success rate drops because of incompatibilities between the different agents. There are two cases. Either an agent cannot incorporate a pattern and therefore creates a new one, or there is more than one possibility in which case patterns proliferate.

Here is an example of the first case. Suppose agent-1 and most of the other agents have the pattern [$\{1,2\},\{3\},\{5\}$] but agent-2 the pattern [$\{1,6\},\{3,4\},5$]]. Suppose [2,4,5] is not possible due to articulatory constraints. Then if agent-1 produces the pattern [2,3,5], agent-2 cannot integrate it because it clashes with the other slot-fillers in his pattern. Agent-2 therefore has to construct a new pattern (when there is space in his memory) and it will take a while before one of the two patterns is pruned to allow pattern coherence with the rest of the group.Here is an example of the second case. Suppose that two agents have both the following patterns [$\{1,2\},3,\{5,7\}$] and [$8,\{3,2\},\{5,6\}$]. Now agent-1 produces a variation: [$9,3,5$]. Agent-2 has two ways to incorporate it. Either by extending pattern-1 which yields [$\{1,2,9\},3,\{5,7\}$] or by extending pattern-2 which yields [$\{9,8\},\{3,2\},\{5,6\}$]. Each extension gives many more additional patterns (which is in se a good thing because it speeds up the growth of the repertoire) but it makes it also more difficult to establish coherence.

Because mutation stops when agents fail in the imitation game, the search process is given time to recover and eventually agents settle on a repertoire of patterns. This has happened in phase 3. Systematicity is now present and success as well as repertoire size move up. In phase 4 a plateau is reached. Due to memory limitations, agents can not store more patterns and given that patterns are already quite complex, it becomes harder and harder to extend them. So the repertoire does not change much anymore and imitation success reaches a stable state.

These simulation results show that a shared repertoire of patterns can indeed emerge in a population with a pattern-based memory. The advantages compared to the earlier solution, where complex sounds were viewed as undifferentiated units, are obvious: (1) To store a complex sound we only need to store the number of slots, the relative timing over these slots, and which individual sound is a possible filler in each

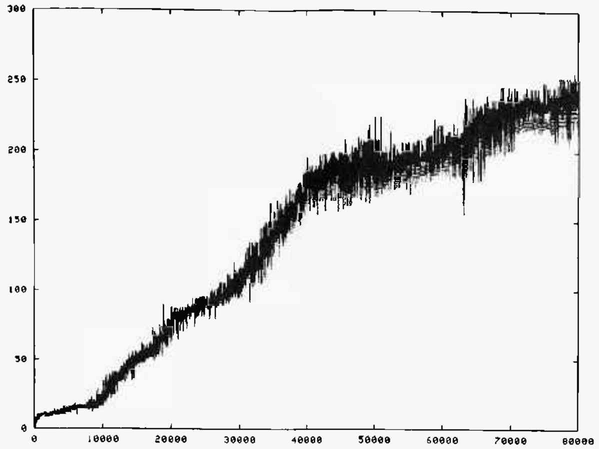


Figure 5: Evolution of the size of the syllable set over time.

slot. So there is an enormous compression of information compared to earlier on. (2) The size of the repertoire now approaches easily (after about 80,000 games) the typical repertoire of human languages, whereas only a dozen stable sound combinations could be handled in the same time frame when no patterns are used. (3) Propagation of new syllables (meaning speed of language acquisition) goes faster, particularly in phase 3 when patterns are already in place. But note that agents must slow down their mutation rate in order to give patterns a chance to stabilise before new ones are introduced.

From sound patterns to categorial constraints

We now turn to a third transition, enabling a full-fledged phonology. The human brain categorises almost anything it is confronted with and then exploits this categorisation to gain in efficiency and reliability. This is also what we postulate as having happened for speech.

Introducing phonological categories

There are many possible sources of categories, given the system described above. For example, certain regions in acoustic space (say high presence of nasal resonance) are systematically associated with certain regions in articulatory space (opening or closing of the nasal cavity), or, the sounds filling a particular slot in a pattern form a natural class, because all of them are viable as fillers of a particular slot and hence satisfy certain articulatory and perceptual constraints. So categories can emerge naturally from the agents' efforts to find variations on patterns and the success they have with particular variations, they do not have to be innate as assumed in abstract generative phonology. We now demonstrate that these natural phonological categories can lead to an additional optimisation in memory usage and in language acquisition.

Let us assume that agents use the natural classes that form in one pattern as a basis for exploring variations in other patterns, both as speaker for producing a new syllable and as hearer for guessing in which pattern they need to incorporate a new syllable. This way the speaker is more likely to produce a new syllable that is viable and the hearer is more likely to integrate the new syllable within existing patterns. This should improve the speed of language acquisition. We call this mechanism *analogy exploitation* because agents construct extensions of one pattern by analogy with another one.

More technically assume that there is a set of patterns $P = \{p_1, ..., pn\}$ where each pattern has a set of slots $p = [s_1, s_2, ...]$ and a set of phonemes associated with each slot. The class associated with slot s_1 in pattern p_1 is denoted as C_{p_1,s_1}. We define a distance metric $\delta(C_1, C_2)$ on the set of phonological classes simply based on the set of common elements.

Assume now that the speaker has an existing pattern p_j but wants to construct a new variation. Rather than selecting a random phoneme from the repertoire of phonemes as the new filler of one of the slots s_k, the speaker searches for the class of sounds C_i that has the shortest distance δ to C_{p_j,s_k}. Then he picks a random member from C_i which was not already in C_{p_j,s_k} and produces the syllable coded by the mutated pattern. For example, the speaker with the patterns given in figure 1 might decide to employ phoneme 1 for constructing a variation of slot3 in pattern2 because $\{1,7,9\}$ have all occurred together as possible slot fillers of slot1 in pattern1. There is no guarantee that the mutation is viable because phonological categories are context-dependent. The possible fillers of a given slot depend on the fillers of the adjacents slots because of co-articulation, and thus the same set of fillers may be unsuited in another context. Nevertheless, the chance that it will be viable is much higher than with a random mutation.

On the side of the hearer, the same mechanism can be used to guess better how a new syllable should be incorporated in the existing repertoire. The hearer computes which patterns cover the new syllable with minimum variation and then use the same distance computation to decide what pattern to change. Most often only one possibility remains and incorporation

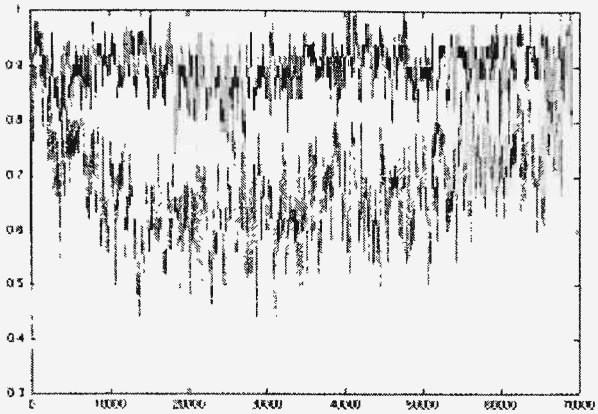


Figure 6: Evolution the percentage of success for 30 agents with analogy exploitation (top) and without (bottom).

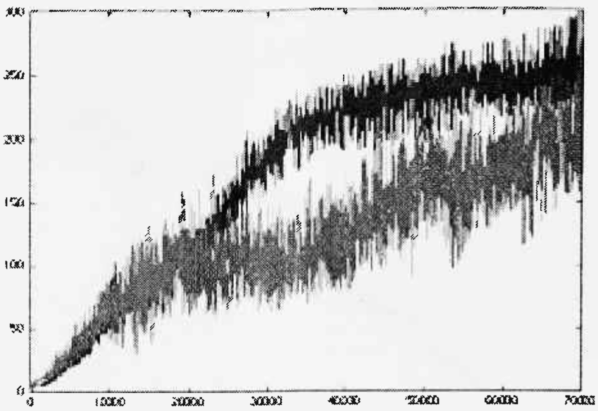


Figure 7: Evolution of number of syllables over time for a group of agents with (top curve) and without (bottom curve) analogy explotation.

becomes a lot less problematic.

Simulation Results

The above mechanisms have again been implemented and subjected to extensive testing. The effect of analogy exploitation can be seen by increasing the mutation rate (from 1 in 100 to 1 in 10 games). Because of this high mutation rate, agents which learn less efficiently will have trouble to acquire the phonology. Recall that mutation stops when success is below 85 % so that agents without analogy exploitation will still catch up but it will take much more time. This is confirmed by the results of experiments displayed in figure 6, which shows the imitation success, and figure 7 which shows the repertoire size. This experiment involves 30 agents and patterns of size 3. We see in figure 6 that the agents which exploit analogy have a consistently higher success rate than those without. Phase 1 is very short. As soon as agents start to make variations on patterns in phase 2, the group without analogy falls below the threshold at which new mutations would occur. A closer inspection of the simulation traces shows that successive incorporation errors lead to incompatibilities that were very difficult to resolve.

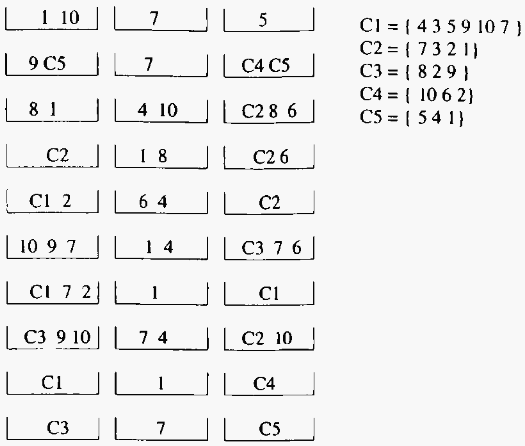
Slower learning can be seen in the evolution of the repertoire size. The group which exploits categorial analogy builds a repertoire much more quickly. The graph shows for each data point (i.e. every 50 games) the highest and the lowest repertoire size in the group. We therefore see that the agents which exploit analogy are also much more coherent throughout.

An example of one set of patterns formed with analogy exploitation is shown in figure 8. Five natural classes have formed. C2 is reused five times. When classes (as opposed to individual phonemenes) are stored we get increased compression although the compression is less dramatic than for the second transition.

Conclusions

The paper has explored a selectionist approach towards the problem how complex syntactic conventions could arise in a population of agents through cultural evolution. New variations are generated and then tested to see whether they are viable from the viewpoint of articulation, distinguishable from the others from the viewpoint of perception, and culturally present in the group. We have introduced three major transitions: from single static sounds to complex undifferentiated sounds, from complex sounds to patterned sounds, and from patterned sounds to categorial constraints. Each transition is caused by recruiting a cognitive device: pattern-based memory as opposed to unit-based memory, and analogical exploitation of the naturally emerging classes, more specifically to restrict the search for patterns that are viable and culturally shared by the speaker or the hearer.

The simulation results obtained so far are extremely encouraging. Nevertheless a lot remains to be done. Most of our future work will be targeted towards adding progressive realism to the simulation. We need to apply the principles discussed in the paper to much more elaborate articulatory synthesisers in which realistic coarticulation effects occur. The recent work of



1 10	7	5
9 C5	7	C4 C5
8 1	4 10	C2 8 6
C2	1 8	C2 6
C1 2	6 4	C2
10 9 7	1 4	C3 7 6
C1 7 2	1	C1
C3 9 10	7 4	C2 10
C1	1	C4
C3	7	C5

C1 = { 4 3 5 9 10 7 }
C2 = { 7 3 2 1 }
C3 = { 8 2 9 }
C4 = { 10 6 2 }
C5 = { 5 4 1 }

Figure 8: Some example patterns in one agent and their slot fillers. Five classes have been formed.

Redford (1999) is an important source of insights on this matter. We will then be able to subject the solutions proposed in this paper to much greater difficulties in acoustic analysis - and we expect to show that categorial constraints can play a major role. Finally we need to couple the mechanism that creates phonological forms to the users of these form, namely lexicon and grammar.

From a broader perspective, we believe that the paper illustrates well how fundamental principles discussed in the context of biological systems are equally present in culturally evolved systems such as language. These principles include self-organisation through a positive feedback loop between use and success and selectionism which combines a process of variation with a process of pruning under natural pressure.

Acknowledgement

We thank Bart de Boer from the VUB AI Laboratory in Brussels for valuable comments on this paper.

References

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

Chomsky, N. and M. Halle (1968) The Sound Pattern of English. Harper Row, New york.

Cook, P.R. (1989) Synthesis of the singing voice using a physically parameterized model of the human vocal tract. Proceedings of the International Computer Music Conference. The MIT Press, Cambridge. pp. 69-72.

De Boer, B. (1999) Investigating the Emergence of Speech Sounds. In: Dean, T. (ed.) Proceedings of IJCAI 99. Morgan Kauffman, San Francisco. pp. 364-369.

Dresher, E. (1992) A learning model for a parametric theory in phonology. In: Levine, R. (ed.) (1992) Formal Grammar: Theory and Implementation. Oxford University Press.

Hardcastle, W.J. and N. Hewlett (eds.) (1999) Coarticulation. Theory, Data and Techniques. Cambridge University Press, Cambridge.

Ladefoged, P. and I. Maddison (1996) The Sounds of the World's Languages. Blackwell Publishers, Oxford.

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.

MacNeilage, P.F. (1998) The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499-548.

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, Mahweh, NJ.

Redford, M.A., C. Chen, and R. Miikkulainen (1998) Modeling the Emergence of Syllable Systems. In: Proceedings of the Twentieth Annual Conference of the Cognitive Science Society. Erlabum Ass. Hillsdale.

Redford, M. A. (1999) An Articulatory Basis for the Syllable. Ph.d. thesis. The University of Texas, Austin.

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

Steels, L. (1997b) The origin of syntax in visually grounded robotic agents. In: Proceedings of IJCAI-97, Morgan Kauffman Pub. Los Angeles.

Steels, L. and R. Brooks (1995) The Artificial Life Route to Artificial Intelligence. Building Embodied Situated Agents. Lawrence Erlbaum, New Haven.

Oudeyer (1999) Experiments in emergent phonetics, Rapport de Stage de 2eme annee de magistere informatique et modelisation, Ecole Normale Superieure de Lyon, Submitted to COGSCI'2000.

Vallee, N. (1994) Systemes vocaliques: de la typologie aux predictions. These. ICP, Grenoble.