Cultural and Biological Evolution of Phonemic Speech

Bart de Boer

Kunstmatige Intelligentie, Rijksuniversiteit Groningen, Grote Kruisstraat 2/1,
9712 TS Groningen, The Netherlands
b.de.boer@ai.rug.nl
http://www.ai.rug.nl/~bart

Abstract. This paper investigates the interaction between cultural evolution and biological evolution in the emergence of phonemic coding in speech. It is observed that our nearest relatives, the primates, use holistic utterances, whereas humans use phonemic utterances. It can therefore be argued that our last common ancestor used holistic utterances and that these must have evolved into phonemic utterances. This involves co-evolution between a repertoire of speech sounds and adaptations to using phonemic speech. The culturally transmitted system of speech sounds influences the fitness of the agents and could conceivably block the transition from holistic to phonemic speech. This paper investigates this transition using a computer model in which agents that can either use holistic or phonemic utterances co-evolve with a lexicon of words. The lexicon is adapted by the speakers to conform to their preferences. It is shown that although the dynamics of the transition are changed, the population still ends up of agents that use phonemic speech.

1 Introduction

All spoken human languages are phonemically coded, that is, they have a large repertoire of words that are built up of a far smaller number of basic building blocks[1]. Thus the words "tea" and "eat" have different meanings, even though they are made up of the same basic sounds. The repertoires of calls of higher primates, on the other hand, are not phonemic. Although their calls are sometimes made up of smaller units such as in the long calls of gibbons, [3] it is not likely that the order of the units influences the meaning of the calls, or that new calls can be created by rearranging the units. Such systems are called *holistic* in this paper.

As our closest evolutionary relatives (Bonobos, Chimpanzees, Gorillas, Orangutans) all use holistic call systems, it can be safely assumed that our last common ancestor also used a holistic call system. At some point in evolution the call system must have made the transition from a holistic to phonemic. This paper addresses an aspect

[1] It is likely that humans use a combination of holistic and phonemic storage. Infants probably store the first words they learn very accurately, and analyze them into building blocks only later *e.g.* [1]. In adult language, too, there are some utterances that have communicative function and are learned, but that fall outside the phonology of the language (called "protosyllabic fossils" in [2]). Examples are utterances such as "pssst", "pfffff", and "tsk tsk". Such utterances are probably stored holistically.

M. Capcarrere et al. (Eds.): ECAL 2005, LNAI 3630, pp. 614 – 623, 2005.

of the question of how this transition can have taken place, and investigates it with the use of a computer model.

Phonemic call systems have a number of advantages over holistic call systems. As they make use of a limited number of discrete building blocks, utterances become more robust. Small errors in pronunciation of a building block will not immediately change it into another building block. This is an example of categorical perception, a phenomenon that is important in the perception of speech [4]. Also, phonemic systems can be made productive: new utterances can be formed by recombining the building blocks in novel ways. Finally, phonemic coding makes it possible to store large repertoires of utterances more compactly. Whereas holistic utterances need to be stored in complete detail, phonemic utterances can be stored in terms of strings of building blocks, while only the building blocks themselves need to be stored in detail. It is this aspect that this paper will focus on.

Although phonemic coding is advantageous for systems with a large number of utterances, holistic systems appear to be preferred for smaller numbers of utterances. This is understandable, as storage complexity and robustness are comparable for small systems, while new utterances can also be created easily in both types. The cognitive complexity of a phonemic system, however, is much higher. It is therefore understandable that a call system would evolve from holistic to phonemic as it grows in size.

This paper does not model the emergence of phonemic coding as such, as is done in for example [5], or the evolution of learning behavior [6], but focuses on the dynamics of the interaction between cultural and biological evolution. The influence of culture is important in the evolution of language *e. g.* [7, 8], but it could be seen as a complicating factor in the transition from a holistic to a phonemic sound system: languages adapt to the abilities of the language users [9] and it can therefore be assumed that a population of holistic learners will shape the language towards holism. As it is assumed that holism is a better strategy for small systems, the system that exists in the population will at first be optimized for holistic learners. When the system becomes bigger, it could become stuck in a state where phonemic coding would *in principle* be more optimal, but where the existing holistic system (as preferred and perpetuated by the holistic learners in the population) causes the fitness of phonemic learners to remain low. This paper investigates the interaction between cultural evolution of a repertoire of sounds and the "biological" evolution of the acquisition strategies in a population of language learners.

2 Phonemic and Holistic Acquisition

In a computer model that investigates the evolution of phonemic acquisition, there must be agents that can both learn a system of speech sounds as a set of holistic motor programs, or as a set of utterances that consist of smaller building blocks and that are thus phonemically coded. In order to implement this, accurate definitions of holistic and phonemic storage are required. In this paper, a holistic system uses only one level of storage. All utterances in the lexicon are stored as exact motor programs. A phonemic system, on the other hand, uses two levels of storage. These are the level of the lexicon and the level of the gestures that make up the building blocks out of which the lexicon is built up. In real language, these could be phonemes, syllables, gestures or

any other realistic primitive. The gestures that make up the building blocks are stored exactly, just as in the holistic system. The lexicon, however stores words as sequences of building blocks, without specifying the details of the building blocks. Learning and storing a pointer to a building block requires less space and effort than learning and storing the actual gestures of the building block. A holistic gesture must be stored with maximum accuracy, as a small change might result in a complete change in meaning. As fewer basic gestures are used in a phonemic system, the margin for error is greater, and storage and learning becomes easier. The two systems are illustrated in figure 1.

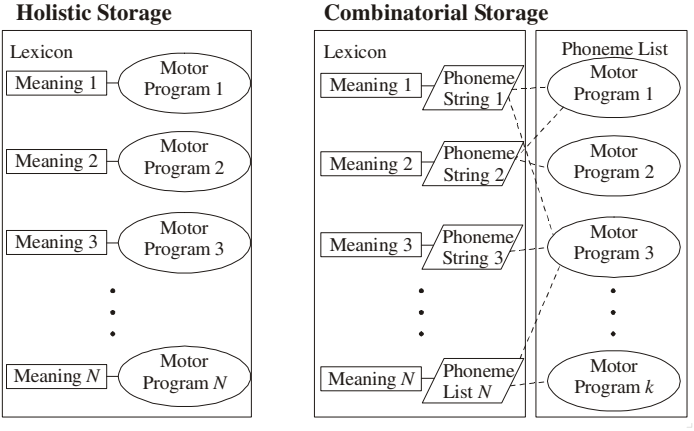

Fig. 1. Diagrams of holistic storage and phonemic storage

If the number of motor programs that are actually used in the lexicon is much smaller than the number of possible motor programs, it can be assumed that the storage of complete motor programs requires more space than storing symbols that represent them. In that case, the lexicon can be stored more compactly using phonemic storage than using holistic storage.

The fitness of agents that use holistic and phonemic coding is determined by a number of factors. One possible factor is the amount of storage required to store a given lexicon. Another factor is the robustness and the communicative success in noisy conditions that can be achieved *e. g.* [10]. A third factor is the amount of cognitive effort that is needed for learning, processing and producing the lexicon. In this paper only the storage requirement is taken into account.

The space s necessary for storing a lexicon L is as follows:

$$s = \begin{cases} \sum_{\forall w_i \in L} \alpha \cdot l_i & \text{, for a holistic system} \\ \alpha \cdot n + \sum_{\forall w_i \in L} \beta \cdot l_i + \gamma & \text{, for a phonemic system} \end{cases}, \tag{1}$$

where α is the number of bits needed to store all information about the motor program of a certain gesture, l_i is the number of gestures in word w_i, n is the number of differ-

ent articulatory gestures (building blocks or "phonemes") used by a phonemically coding agent, β is the number of bits needed to specify a phoneme in a word and γ the overhead needed for a phonemically coded system. From these equations it is clear that holistic coding is more efficient for a lexicon that has only words consisting of a single gesture. Phonemic coding is more efficient if there are many words re-using the same gestures in the lexicon. The exact parameter values define the transition point. It has not been attempted to estimate realistic values for the parameters. Our knowledge of how speech is stored in the brain is insufficient for this. Nevertheless, there are a few relations between the parameters that can be determined from first principles. It must be true that $\alpha \gg \beta$, as specifying a gesture exactly is more complex than referring to it. It must also be true that $\beta \equiv \log_2 n$. The more phonemes there are, the more bits are needed to distinguish them, and information theory [11] teaches that this number is proportional to the logarithm of the number of phonemes. Furthermore, it is impossible to use all potential articulatory gestures as distinctive speech sounds (phonemes). In order to preserve acoustic distinctiveness with a margin of error there must be unused acoustic and articulatory space between different gestures. This means that there are fewer possible phonemes than possible gestures. This can be formulated in the equation: $n_{\max} \ll 2^{\alpha}$, where $n_{\max}$ is the maximal number of phonemes, and 2^{α} is the maximum number of possible gestures, as this is the number of different strings of α bits (if there were more gestures, more bits would be needed to distinguish them).

The above considerations imply that a lexicon with a sufficiently large number of words always needs to contain words that consist of multiple gestures. Therefore, for sufficiently large lexicons, phonemic coding will be more efficient. If the lexicon is small, on the other hand, it can consist of single-gesture words without impeding distinctiveness. Therefore holistic coding can be more efficient.

At some intermediate size, a transition from optimality of holistic systems to optimality of phonemic systems must occur. At what size the transition will actually take place is not just determined by the parameters (which can be considered genetically determined factors) but also by what system is already in use in a population, i.e. cultural factors. Holistic learners will prefer a system with many different gestures and short words, while phonemic learners prefer a system with long words and few different gestures. Through self-organization, the sound system in the population will tend to adapt to the preferences of the majority of the population. This will have effects on the dynamics in a system where both the learned system of speech sounds and the learners themselves change.

3 The Model

A computer model has been implemented to investigate the dynamics of a system that combines genetic evolution of learners with cultural evolution of a repertoire of speech sounds. There are two kinds of agents in the model: holistic learners and phonemic learners. This is a simplification, because humans probably use both strategies, but this makes it much easier to understand the dynamics of the model. As there are only two types of agents, the population could have been modeled as a set of one-bit genomes. Instead, it has been decided to model only the fraction of agents that learn

holistically, p_h and the fraction that learns phonemically, p_p. Although this way of modeling a population is perhaps unusual in artificial life, where one generally prefers to model complete agents, it is an old tradition in theoretical biology [12]. Because there are only two agent types, it follows that $p_h + p_p = 1$. As all agents within each group are identical, fitness can be associated with the group instead of with the individuals. The fitness of the two groups is indicated with f_h and f_p, respectively and they are used to calculate the fractions of each type of agent in the next generation. There is also the probability μ of one type of agent mutating into the other type of agent (set to 0.1 in the simulations presented here). The equations for calculating the proportion of agents in the next generation are as follows:

$$\begin{aligned} p_{h,t+1} &\leftarrow \nu\left(f_{h,t} \cdot p_{h,t} + \mu \cdot f_{p,t} \cdot p_{p,t}\right) \\ p_{p,t+1} &\leftarrow \nu\left(f_{p,t} \cdot p_{p,t} + \mu \cdot f_{h,t} \cdot p_{h,t}\right) \end{aligned}, \tag{2}$$

where ν is a factor that causes $p_{h,t+1}$ and $p_{p,t+1}$ to sum to one.

All agents in the population, both holistic and phonemic learners, share the same lexicon. The fitness of each group is determined by the number of bits needed to store this lexicon. The lexicon consists of a list of words that each in turn consist of one or more basic gestures. These basic gestures can be represented by symbols, and phonemic learners use them as their building blocks (phonemes). Equation 1 is used for calculating the number of bits needed to store a repertoire. An example of a repertoire and the number of bits needed to store it is given in figure 2. The values of the parameters used to calculate these numbers are as follows: number of bits for a gesture (α): 10, penalty for using a phonemic system (γ): 30 bits, number of bits used per phoneme (β): $\log_2 n$. These same parameters were used in all the simulations as well. Note that it is possible to optimize the lexicon used in the example for both holistic and phonemic learners. For holistic learners, a system using the same gestures and having the same number of words, but needing only 110 bits of storage can be constructed (by changing *ao* into *o* and *aea* into *ea,* for example). Phonemic coding could also be substantially more efficient, for example by using two one-phoneme, four two-phoneme and two three-phoneme words. In that case only two different phonemes would be needed, resulting in a size of 66 bits. The resulting lexicons are shown on the right side of figure 2. It is clear that for lexicons of the same size, holistic and phonemic learners prefer very different words.

Given the number of bits s_p and s_h for phonemic and holistic learners, respectively the fitnesses are calculated as follows:

$$\begin{aligned} f_p &= 1 - \frac{s_p}{s_p + s_h} \\ f_h &= 1 - \frac{s_h}{s_p + s_h} \end{aligned}. \tag{3}$$

Note that fitness is relative: the fitness of one group of agents depends on the fitness of the other group. Thus the two groups *co-evolve*.

Each generation, the lexicon can be modified. A new word can be added with a certain probability and the words in the lexicon are modified in correspondence with the preferences of the agents in the population. The new word that is added is the

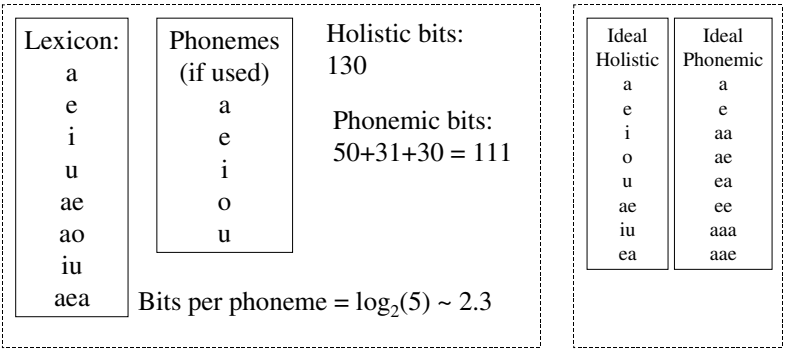

Fig. 2. Example of a lexicon and the number of bits needed for holistic and phonemic storage. For parameter values, see the text.

shortest word that is not already present. Addition is a holistic process, in the sense that new articulatory gestures can be created. It was decided to implement addition as a holistic process, because phonemic addition would not as readily add new phonemes to a growing repertoire, and because in the stage before transition, the majority of the population is expected to consist of holistic learners. A possible variant of the model would be to make the type of addition used depend on the proportion of holistic and phonemic agents in the population. As words can sometimes be removed from the lexicon the added word can be shorter than the longest word in the lexicon. Words are added with a probability of 10% per generation.

The lexicon can also be modified to better suit either type of agent. To suit holistic agents (who dislike long words) the longest word is removed from the lexicon and replaced with an unused shorter word, if possible. New articulatory gestures can be introduced as a side effect. To suit phonemic agents, first the phoneme that is least often used is found, and then the word in which it occurs most frequently. It is then attempted to replace this word with the shortest word that is build up of the phonemes already present in the lexicon. This can cause phonemes to disappear from the lexicon and average word length to increase. The first process puts pressure on the lexicon for shorter words and more phonemes, while the second process puts pressure on the lexicon for longer words and fewer phonemes.

If culture is turned on in the simulation presented here, two words can be modified per generation. No modification takes place if there is no culture. For each modification either a holistic or a phonemic agent is selected with probabilities that are proportional to their abundance in the population. Thus, if there are many holistic agents, the lexicon is pushed towards holism. If there are many phonemic agents, it is pushed towards phonemic coding.

4 Results

In the experiments, the population is initialized with 50% holistic and 50% phonemic agents. The lexicon is initialized with a single word (consisting of a single gesture). The maximum number of different gestures (n_{max}) is 16. The rest of the parameters are

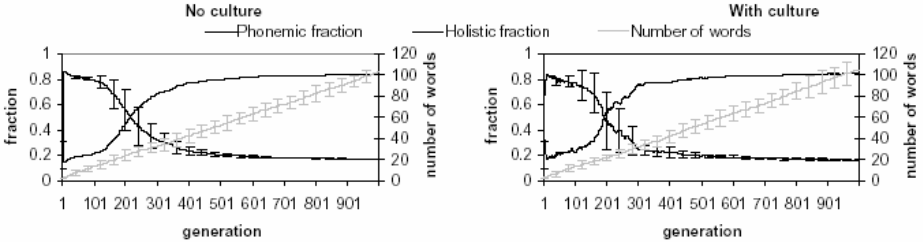

Fig. 3. Comparison of average behavior (over 10 runs) of a population without culture (left graph) and a population with culture (right graph). As the fractions of phonemic and holistic fractions are symmetrical, error bars showing standard deviation are only shown for the holistic fraction.

as described above. The result of running the model without and with cultural influences is shown in figure 3. As can be seen from this figure, the proportion of holistic agents rapidly rises in the beginning. Holistic agents have higher fitness for small lexicons than phonemic agents (as would be expected). The population then remains almost exclusively holistic for a while (a minority of phonemic agents remains present because of mutation). After a critical number of words is reached, the population makes a transition from a holistic majority to a phonemic majority.

At first sight, there seems to be little difference in average behavior between systems with culture and without culture. However, as can be observed in the graph, the error bars in the graph for populations with culture are much larger. Apparently there is a difference between the two cases, but it is masked by the averaging procedure.

The difference is caused by the fact that the transition is much faster for populations with culture than for populations without. This is illustrated in figure 4 which shows typical runs for systems without and with culture. For faster transitions, the standard deviation will be calculated over runs in which some of the populations are still in the holistic state and others in the phonemic state, hence increasing the standard deviation. But this only reflects the fact that the standard deviation is calculated over a distribution with two peaks, not that the peaks themselves are broader.

A comparison of the number of generations needed for the fraction of holistic agents to change from above 60% to below 40% confirms this. This is 52.3 generations ($\sigma = 14.6$) for the population without culture and 12.1 generations ($\sigma = 5.6$) for the population with culture. There appears to be no significant difference for the time at which the drop takes place; this happens after 191 ($\sigma = 42.6$) and 194 ($\sigma = 49.7$) generations, respectively. There is a remarkably large variation in the number of generations to the transition, but this is due to the fact that the increase of the number of words is a random process. The large variation disappears when one looks at the number of words at which the transition takes place. The number of words at which the drop starts (the 60% holistic learners threshold is crossed) is 22.4 ($\sigma = 0.52$) for populations without culture and 21.4 ($\sigma = 1.07$) for systems with culture. Although this is a significant difference, it is probably caused by the fact that populations with culture go through the transition faster than populations without culture and it probably does not reflect a real difference in the size of the lexicon at which transition starts. Systems probably start to transition when the number of words exceeds the

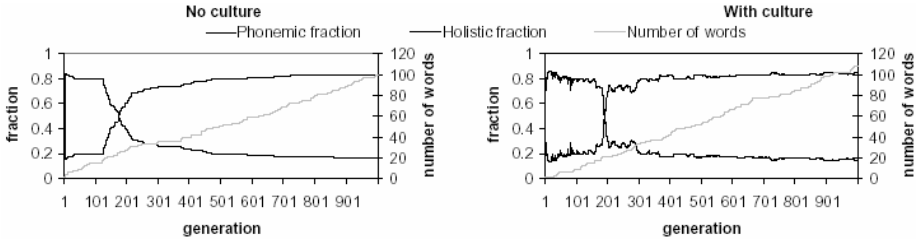

Fig. 4. Typical runs from a system without culture (left graph) and a system with culture (right graph). Note the faster transition in the graph with culture.

number of possible articulatory gestures. This is borne out by inspection of the data sets and confirmed for systems with 32 instead of 16 possible different gestures. The number of combinatorial agents starts to rise once complex utterances become necessary. Even systems that have been optimized for holistic learners then become more efficient for phonemic learners.

These results are robust for changes of parameters. As has been said above, changing the number of possible gestures does not affect the qualitative behavior. Changing the number of modifications that are made to the repertoire does not appear to affect the results. Changing the mutation rate to a lower value (0.03) causes agents of the unfavored type to become rarer. As it takes slightly longer for the number of agents to rise when the transition starts, this causes transitions to occur slightly later (at 24–25 words) An example is given in figure 5.

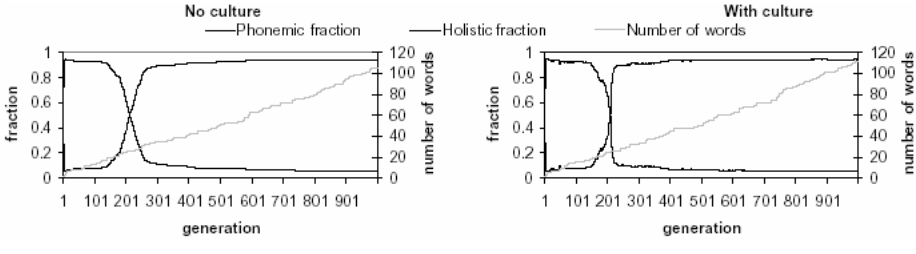

Fig. 5. Typical runs from a system with low mutation rate (m=0.03). Note the similarity with the transition in the previous figure.

5 Discussion and Conclusion

In the experiments presented here, holistic and phonemic learners had to compete. Their fitness depended on the amount of storage that was needed for storing a lexicon that evolved (culturally) together with the agents. Both types of agents changed the lexicon such that it would be easier for them to learn, causing the lexicon to tend towards optimality for the agents that had the majority in the population. It could be imagined that these cultural dynamics could cause the system to remain stuck in a local optimum, such that it would not make the transition from the (initially optimal) holistic lexicon to the

(ultimately optimal) phonemic lexicon. This was not observed in the experiments for different parameter settings. It can therefore be tentatively concluded that cultural inertia does not greatly influence the transition from holistic to phonemic speech.

The dynamics of a population with culture was observed to be different than that from a population without culture, however. It was observed that a population can change much more abruptly from holistic to phonemic language use if there is co-evolution of the culturally transmitted system of speech sounds with the genetically evolving language learners. This can be explained by the fact that once the proportion of phonemic learners starts to increase, the lexicon will also be changed to become more optimal for the phonemic agents. This increases the fitness of the phonemic agents, thus accelerating the transition. So although culture might at first be an obstacle to the transition, in the end it accelerates it.

These observations illustrate that the evolution of speech (and language) cannot be seen as either purely genetic or purely cultural evolution. Both mechanisms must be taken into account. They also confirm that phonemically coded systems win over holistically coded systems, at least as far as storage is concerned. They win, even though at first there is a cultural evolution towards systems that are more learnable for holistic learners.

The model that was used is admittedly simplistic, and only a limited number of experiments was performed. The model was made so simple in order to create a very basic model that nevertheless has interaction between the evolution of the acquisition of complex speech and the (cultural) evolution of the speech sounds themselves. There are many ways in which this work can be extended: a mathematical analysis of the dynamics would seem to be possible. Also, the influence of the different parameters can be investigated, it could be investigated at which point the transition from holistic to phonemic coding takes place exactly, different variants on the optimization procedures and the addition of new words could be tried out and many other small variants.

More interesting, however, is to strive for more realism in the simulation. As the model is about the evolution of acquisition, it is important to try to model a population of agents that really acquire the system of speech sounds. The acquisition mechanism should have parameters that make it exploit phonemic structure to a higher or lower degree, and these parameters should be able to evolve. The time it takes to acquire a repertoire of speech sounds and the accuracy with which this happens could be taken into account in the fitness function. A next step could then be to get rid of the global lexicon, and have them be emergent in the population, just as the sound systems are emergent in the population in [13, 14]. Also, more realistic constraints on production and perception could be added. Such a model would already be quite realistic, but also have much more complicated and hard-to-understand dynamics. The model as presented in this paper gives a basis for understanding such dynamics.

Another important research topic would be finding independent evidence with which to compare the results of such a computer simulation. This can be evidence from language acquisition, evidence from the fossil record, or evidence from animal call systems. Perhaps the study of call systems from closely related primate species (the different species of gibbon, [15] for example) can provide insight into the circumstances under which a holistic call system can change into a phonemic call system.

The understanding of the dynamics of the evolution of speech is a crucial piece of the puzzle that cannot be found either by studying animals or by studying the fossil record. The interaction between the evolution of a cultural repertoire of speech sounds and the physical adaptations for processing, producing and perceiving them cause the evolution of speech to have very complex dynamics. Computer models can provide insights in these dynamics. The computer model presented here is an attempt to provide insight in the interaction between cultural and genetic evolution, and it is hoped that it can be used as an inspiration for further research.

References

1. Houston, D. M. & Jusczyk, P. W.: The role of talker-specific information in word segmentation by infants. Journal of Experimental Psychology: Human Perception and Performance 26: (2000) 1570–1582
2. Jackendoff, R.: Foundations of language. Oxford University Press, Oxford (2002)
3. Mitani, J. C. & Marler, P.: A phonological analysis of male gibbon singing behavior. Behaviour 109: (1989) 20–45
4. Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M. & Gerstman, L. J.: Some Experiments on the Perception of Synthetic Speech Sounds. Journal of the Acoustical Society of America 24: (1952) 597–606
5. Oudeyer, P.-Y.: Phonemic coding might be a result of sensory-motor coupling dynamics. In: Hallam, J. (ed.): Proceedings of the International conference on the simulation of adaptive behavior (SAB). MIT Press, Edinburgh (2002) 406–416
6. Hurford, J.: The evolution of the critical period for language acquisition. Cognition 40: (1991) 159–201
7. Steels, L.: Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation. In: Hurford, J. R., Michael, S.-K. & Knight, C. (eds.): Approaches to the Evolution of Language. Cambridge University Press, Cambridge (1998) 384–404
8. Smith, K., Kirby, S. & Brighton, H.: Iterated Learning: a framework for the emergence of language. Artificial Life 9: (2003) 371–386
9. Zuidema, W.: How the poverty of the stimulus solves the poverty of the stimulus. In: Becker, S., Thrun, S. & Obermayer, K. (eds.): Advances in Neural Information Processing Systems 15. MIT Press, Cambridge, MA (2003) 51–58
10. Nowak, M. A. & Krakauer, D.: The evolution of language. Proceedings of the National Academy of Sciences 96: (1999) 8028–8033
11. Shannon, C. E.: A mathematical theory of communication. The Bell system technical journal 27: (1948) 379–423,623–656
12. Maynard Smith, J.: Models in Ecology. Cambridge University Press, Cambridge (1974)
13. De Boer, B.: The origins of vowel systems. Oxford University Press, Oxford (2001)
14. De Boer, B.: Self organization in vowel systems. Journal of Phonetics 28: (2000) 441–465
15. Geissmann: Duet-splitting and the evolution of gibbon songs. Biological Reviews of the Cambridge Philosophical Society 77: (2002) 57–76