Evolution and self-organisation in vowel systems*

Bart de Boer
Vrije Universiteit Brussel

This paper describes computer simulations that investigate the role of self-organisation in explaining the universals of human vowel systems. It has been observed that human vowel systems show remarkable regularities, and that these regularities optimise acoustic distinctiveness and are therefore adaptive for good communication. Traditionally, universals have been explained as the result of innate properties of the human language faculty, and therefore need an evolutionary explanation. In this paper it is argued that the regularities emerge as the result of self-organisation in a population and therefore need not be the result of biological evolution.

The hypothesis is investigated with two different computer simulations that are based on a population of agents that try to imitate each other as well as possible. Each agent can produce and perceive vowels in a human-like way and stores vowels as articulatory and acoustic prototypes. The aim of the agents is to imitate each other as well as possible.

It will be shown that successful repertoires of vowels emerge that show the same regularities as human vowel systems.

1. Introduction

This paper explores the possibility that next to natural evolution, self-organisation has played an important role in the way human languages, and more specifically their sound systems, are formed. Languages are highly complex phenomena. They have huge vocabularies, often have complex syntax and morphology and elaborate sound systems. If one looks at the different languages of the world, one finds a bewildering diversity in all these aspects. On the other hand, there are also a number of regularities to be observed. A part of the

Evolution of Communication 3:1 (1999), 79–102. ISSN 1387–5337

study of human language has focused on these regularities, which are also called universals. Knowing the universals of human language helps to establish what a theory of the evolution of human language has to explain. However, natural evolution is a rather slow process, and it appears that human language has emerged in a relatively short time (either since the appearance of Homo Erectus two million years ago or since the appearance of modern Homo Sapiens, ~300 000 years ago, see e.g. Hurford et al., 1998). It is therefore desirable to explain as many properties of human language by other means than evolution. The hypothesis on which the work in this paper is based is that self-organisation is one of these other means.

In this paper, the term self-organisation (Nicolis & Prigogine, 1977) is meant to refer to the emergence of global order from local interactions in complex systems. A complex system in this definition is a system that consists of a number of interacting simpler elements. The global behaviour cannot be described as a simple sum of the behaviour of the constituent elements. Global order is an order that is extended over a large number of the constituent elements of the system, so that in principle the state of the total system can be described by fewer variables than there are elements in the system. Local interactions are interactions in which only a few constituent elements are involved. This excludes either global control of the system, while also hierarchical organisation of the system is excluded.

An excellent example of a self-organising system is the honeycomb. A structure on a global scale (the hexagonal grid of cells) emerges out of local interactions between individual bees, without global control. The hexagonal grid appears because bees that are adding adjacent cells to the grid are approximately the same size and compress the wax of which the cell's walls are constructed with approximately the same force. So although there are many different ways in which one can place cells in a plane, the resulting grid is very regular and can be described by only a few variables (the size of the cells and the angle of the tiling, for example).

Self-organisation in language can take place on at least two levels, that of the individual and that of the population. On the level of the individual, self-organisation can play a role in the way linguistic knowledge is organised in the brain. It is known from research into vision that self-organisation plays a role in the way the left eye and the right eye are mapped to adjacent columns of the visual cortex (Erwin et al., 1995). Similar processes might occur in the organisation of linguistic knowledge in the brain. However, this paper is not concerned with self-organisation on this level. The self-organisation that is investigated in

this paper is on the level of the population. In this type of self-organisation, universal properties of the language emerge as the result of constraints on the individuals and their interactions. This kind of self-organisation provides a link between language as the knowledge of an individual and language as a population phenomenon.

The idea that self-organisation plays a role in language is not new. However, most work has either been purely theoretical or philosophical (Petitot-Cocorda, 1985; Wildgen, 1990) or focused on self-organisation in individuals (Lindblom et al., 1984). Only with the advent of cheap computing power have full-scale investigations of self-organisation in language begun (Steels, 1997).

It will be shown in this paper that self-organisation can provide an alternative to having to find an evolutionary explanation for each and every language universal. Clearly some adaptations for language are the result of natural selection. An excellent example of this is the human vocal tract. Only an explanation in terms of evolution makes sense in explaining it. However, for many other properties of language, an evolutionary explanation is much more complicated. Why for example, would mid front vowels (such as [e]) occur more often than mid back vowels (such as [o]) in human languages? And where would such a universal tendency be coded, both in the genome and in the brain? It would be much preferable if one could find a non-evolutionary explanation for such phenomena.

The research described in this paper focuses on vowel systems. It was decided to investigate sound systems instead of other parts of human language, because these are the most concrete and therefore the most verifiable part of human language. Within human sound systems it was decided to focus on vowels for two reasons. First, there is a lot of data about the universal tendencies of human vowel systems (see the next section) and about the way human vowels are produced and perceived. It is therefore easy to check the predictions of a theory or a model against human data. Secondly, it is easy and computationally efficient to build computer models of human perception and production of vowels. More complex utterances are much more difficult as will be explained in the section on future work.

The research depends heavily on the use of computer simulations. The emergence of order through self-organisation in a population is a phenomenon that by its very nature is very hard to predict a priori. It is therefore necessary to test the theories and hypothesis by actually implementing them. For this purpose simplified models of perception, production and learning have to be made. The biggest problem of implementation is finding the right simplifications. Too much simplification and the results will no longer be realistic, too little simplification and the model will be computationally infeasible. In the section on the simulations, the design decisions that have been made will be presented and defended.

2. Universals of vowel systems

Vowels are defined as voiced oral utterances in which there is an unrestricted airflow and in which the articulators are static. Although this is only a very limited subset of the possible articulations that can be made by the human vocal tract, there is still a wide variety of sounds that can be made in this way. The three main articulatory parameters that determine vowel timbre are tongue height, (the distance between the highest point of the tongue and the palate) tongue position, (the position in the front-back dimension of the highest point of the tongue) and lip rounding. In human languages, many other possible variations of the basic vowel are possible, such as nasalisation, pharyngealisation or lengthening. However, these factors will not be considered in this paper. Ladefoged and Maddieson (1996: Chapter 9) find that in a single human language, it is possible to have five possible degrees of tongue height, three degrees of tongue position and three degrees of lip rounding. However, not all these combinations occur in a single human language. In the UPSID$_{451}$, the UCLA Phonological Segment Inventory Database with 451 languages, (Maddieson, 1984; Maddieson & Precoda, 1990) the languages with most contrasts in basic vowel timbre are Norwegian and German, both with 15 different vowel timbres (although in both languages there is also a length distinction).

As there are many more vowels than there are vowel symbols in the alphabet — one should take care not to confuse the two; especially in English both the same vowel sound is written with different symbols and the same symbol is used for writing different vowel sounds — they are usually written with symbols from the international phonetic alphabet (IPA) and this is the convention that is used in this paper as well. There is a further important convention that is used in writing down sound systems of human languages. When one writes down the actual realisation of a speech sound, one makes a phonetic transcription, and this is written in square brackets, e.g. [i] (the vowel sound of "beet"). However, if one writes down the category to which a sound belongs in a certain language, one makes a phonological transcription and the vowel will be written between slashes: /i/. This is important, because often vowels are subject to change due to co-articulation, so that their actual phonetic realisation can differ, while they still belong to the same phonological category. As the phonetic realisation of speech sounds of a language differs considerably from speaker to speaker, inventories of speech sounds will be written between slashes in this paper. Examples of utterances will be written between square brackets.

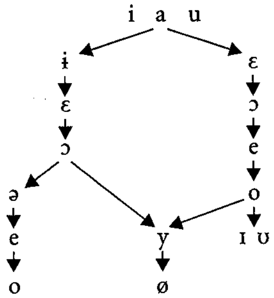

Figure 1. Vowel system hierarchy according to Crothers (1978).

Still, although humans are able to produce and distinguish many different vowels in principle, it has been found that the vowel systems that actually occur in human languages are quite regular and predictable. By far the most frequently occurring vowel system is the one consisting of the five vowels: /i/, /e/, /a/, /o/ and /u/. More generally, certain vowels occur much more often than others. The vowels /i/, /a/ and /u/ occur in almost all languages (in 87%, 87% and 82%, of the languages in UPSID$_{451}$, respectively) while the some vowels, such as the /ʌ/ (the vowel of English "but") occur only rarely (in 2% of the languages in UPSID$_{451}$). In general, peripheral vowels (vowels that are articulated with the articulators in one of their extreme positions) tend to be preferred over non-peripheral vowels.

Also, many symmetries occur. For example, whenever a language contains a back rounded vowel of a given height (such as [o] which occurs in 29% of the languages in UPSID$_{451}$), it is much more likely to contain the front unrounded vowel of the same height (in this case [e] which occurs in 27% of the languages of UPSID$_{451}$, but in 80% of the languages which also contain [o]) than would be expected from the a priori frequency of that vowel. A number of universal tendencies of vowel systems have been found (Crothers, 1978; Vallée, 1994, Schwartz et al., 1997a). Crothers (1978) summarised the universals he derived for simple vowels in the following diagram (Figure 1). This diagram indicates the sequence in which vowels tend to appear in larger and larger vowel systems.

Almost all languages with three vowels have /i/, /a/ and /u/. Systems of four vowels tend to have /ɛ/ or /ɨ/ in addition etc.

Traditionally, the shapes and universal tendencies of human vowel systems have been explained as the result of distinctive features and their markedness (Jakobson & Halle, 1956; Chomsky & Halle, 1968). However, distinctive feature theory requires that there must be innate cognitive structures that in turn have to be explained by biological evolution. In this paper, it will be shown that for the universal tendencies of vowel systems, such innate properties are not necessary *in principle.*

It appears that the most frequently occurring human vowel systems are maximally dispersed through the available acoustic space. Liljencrants and Lindblom (1972) have shown that when one optimises simple vowel systems for maximal acoustic distance between the different vowels in the system, one obtains vowel systems that occur frequently in human languages. Their original work has been improved upon by others (e.g. Schwartz et al., 1997b). This is a nice functional explanation of the universals of vowel systems, and it explains very well *why* human vowel systems are the way the are, but not *how* they have become this way.

Human vowel systems appear to be optimised for acoustic distinctiveness, but it is not clear who is doing the optimisation. In fact, children that learn a new language do not optimise the sound system of the language they learn. They gradually approach the sound system that is used by the people in their environment as closely as possible. In this way it is possible that within a language there can be recognisable regional and social variations of the sound system. Whereas the meaning of what is said is clear, subtle difference in pronunciation make it possible to identify a speaker's region of origin and sometimes his or her social class. The hypothesis that is tested in this paper is that the optimisation is the effect of self-organisation in the population of language users.

The first to investigate universals of vowel systems using computer models of populations of agents was Glotin (Glotin, 1995; Glotin & Laboissière, 1996). He made a simulation of a population of agents that each started with a repertoire of a predetermined number of randomly initialised vowels with both an acoustic and an articulatory representation. These agents would interact in pairs. One would pick a vowel from its repertoire and the other would move its closest vowel closer to the one it had heard, while moving away its other vowels. The effort of moving vowels would be calculated for each agent. The less effort an agent had had to do, the more fit it was. The fittest agents were allowed to create offspring, whose vowel systems were a mixture of their parents' vowel systems. Glotin has shown that for small numbers of agents and small numbers of vowels, coherent and more or less realistic vowel systems would emerge, although sometimes the agents would not be able to develop a coherent vowel system.

Glotin's simulation suffers from two drawbacks: first of all, because of the elaborate articulatory model he used, his simulations were extremely slow, and only experiments with small populations could be done. But more importantly, his simulation contains a number of elements that interact, and that make it difficult to determine which effects are caused by which aspect of the simulation. There is an element of self-organisation caused by the interactions between the agents, but there is also an element of pure optimisation, because the agents move their vowels away from each other, while there is also a genetic component based on the amount of shifting an agent does with its vowel system. Whatever its drawbacks, the work described in this paper is based in a large part on elements of the work of Glotin.

Glotin's simulation has been elaborated upon by Berrah (1998). Berrah removed the articulatory representation from Glotin's simulation and worked in acoustic space only. He also removed the genetic component from the simulation. The optimisation performed by the agents by pushing their vowels away from each other was kept, however. For this reason Berrah's simulation is in the end similar to Liljencrants and Lindblom's optimising simulation, as Berrah has remarked himself in his experiments. The merit of Berrah's work is that he has extended the simulations to accommodate extra parameters, such as length or nasalisation, but this falls outside the scope of this paper.

This paper is concerned with the investigation of whether a coherent and realistic vowel system can emerge in a population of agents that can only have local (one-to-one) interactions and that cannot do global optimisations of their vowel systems. In this respect the phenomenon under investigation is exactly self-organisation according to the definition given above.

3. The simulations

The work described in this paper falls within the framework of Steels' (1996, 1997) research into the origins of language. One of the bases of this research is that the proposed systems and theories should work in the "real world". Although within Steels' framework, research has been done on real robots (Steels, 1998a, b; Vogt, 1998), the research described here has been done in simulation, in order to make it computationally feasible. However, the agents use both an acoustic and an articulatory representation of their vowels in order to stay as close as possible to the kinds of constraints that humans have. Also, the mechanisms behind the simulation have been kept as simple as possible in order to be able to interpret the results more easily. Neither genetic mechanisms nor neural networks have been used. Finally the agents can only make biologically plausible updates to their vowel systems. They cannot look into each other's head and they can only use local information. This consists of the feedback received from one imitation game, as well as information about only one vowel. Global optimisation is not possible in this way.

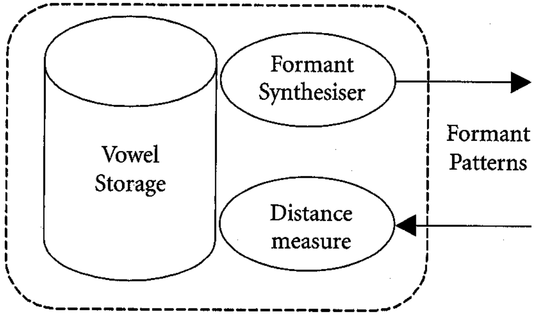

Figure 2. Simplified agent architecture

The simulation is based on a population of agents. Each of these agents can produce and perceive vowels in a realistic, human-like way. The agents' goal in life is to imitate each other as well as possible. For this purpose they have an articulatory synthesiser with which they can produce formant patterns given an articulatory input. They also have an acoustic model with which they can calculate distances between formant patterns that correspond to the distances as they are perceived by humans. Finally they have a storage that is based on storing acoustic and articulatory prototypes of vowels they use. Each of these will be described in some detail below.

The agents interact by playing imitation games with each other. The idea that the exchange of linguistic information is like a game is originally from Wittgenstein (1967) who also coined the term "language game." The idea of using language games in computer simulations has first been elaborated by Steels (1996). In a language game in general, there are first of all the participants. Usually there are two participants and these are called the speaker and the hearer, but in the imitation game both participants both speak and listen, so the terms initiator and imitator are used. Secondly there are the rules of the interaction. These determine who speaks and who listens at each moment and what kind of linguistic and extra-linguistic information can be exchanged. Finally, there are the ways in which each agent reacts to the linguistic and extra-linguistic information it receives.

Two types of imitation game are described in this paper. The first is the original imitation game, with which most of the results that will be described in this paper were obtained. This game has already been described in great detail elsewhere (de Boer 2000) so it will only be described in general terms. The second type of imitation game is in fact a simplification of the original imitation game. In this game fewer assumptions are made of what the agents can do, but a disadvantage is that the number of vowels has to be fixed beforehand. However, the second type of game seems at the moment to be more promising if one wants to extend the imitation game to more complex utterances.

3.1 The articulatory model

But first the agent architecture will be described in some more detail. The articulatory synthesiser is a very much simplified model of human articulation of vowels and is based on three observations. Firstly the articulation of a basic vowel can be described by three parameters, tongue position, tongue height and lip rounding. Secondly the acoustic signal of a vowel can be described by three or four numbers that represent the frequency of peaks in its acoustic spectrum. Finally there is a more or less smooth relation between the frequency of these peaks and the acoustic parameters. Therefore the articulatory synthesiser could be based on an interpolation of the relation between the articulatory position and the acoustic signal of 18 basic vowel signals, resulting in a function that can be evaluated very quickly.

The three articulatory parameters correspond roughly to possible movements of the tongue body and the lips. Although both are controlled by a large number of muscles, phoneticians normally abstract away from these and assume that vowel articulations are controlled by the distance of the highest point of the tongue and the palate, (tongue height) the position of this highest point in the front-back dimension (tongue position) and the rounding of the lips. Of course, vowels in human languages are often articulated with extra features: nasalisation, pharyngealisation, retroflexion etcetera. However, these are generally considered to modify only the timbre of the basic vowels.

In the synthesiser, the value for tongue height was arbitrarily defined to be 0 for the lowest vowels on which the interpolation was based and 1 for the highest vowels. Similarly tongue position was defined to be 0 for the most fronted vowels, and 1 for the most back vowels, while lip rounding was defined to be 0 for unrounded vowels and 1 for rounded vowels. For each of these eight extreme values there was a data point, while ten extra data points for vowels in between the extremes whose values for tongue position and/or tongue height were defined to be 0.5 were also taken into account.

Table 1. Synthesiser equations

$$
\begin{aligned}
F_1 &= ((-392+392r)\ h^2 + (596-668r)\ h + (-146+166r))p^2 + \\
&\quad ((348-348r)\ h^2 + (-494+606r)\ h + (141-175r))p + \\
&\quad ((340-72r)\ h^2 + (-796+108r)\ h + (708-38r)) \\
F_2 &= ((-1200+1208r)\ h^2 + (1320-1328r)\ h + (118-158r))p^2 + \\
&\quad ((1864-1488r)\ h^2 + (-2644+1510r)\ h + (-561+221r))p + \\
&\quad ((-670+490r)\ h^2 + (1355-697r)\ h + (1517-117r)) \\
F_3 &= ((604-604r)\ h^2 + (1038-1178r)\ h + (246+566r))p^2 + \\
&\quad ((-1150+1262r)\ h^2 + (-1443+1313r)\ h + (-317-483r))p + \\
&\quad ((1130-836r)\ h^2\ (-315+44r)\ h + (2427-127r)) \\
F_4 &= ((-1120+16r)\ h^2 + (1696-180r)\ h + (500+522r))p^2 + \\
&\quad ((-140+240r)\ h^2 + (-578+214r)\ h + (-692-419r))p + \\
&\quad ((1480-602r)\ h^2 + (-1220+289r)\ h + (3678-178r))
\end{aligned}
$$

For each of these data points the values of the first four formants were taken from (Vallée, 1994). A *formant* is a peak in the acoustic power spectrum of a vowel. Such a peak corresponds to a resonance that the vocal tract has when the tongue and lips are put in the position to articulate that vowel. The positions of these resonances (and in a lesser degree their bandwidth) determine the acoustical quality of the vowel. Notation of formants is done with capital *F* with the number of the formant peak as subscript. The second formant, for example is written as F_2. Based on Ladefoged's (Ladefoged, 1981, ch. 8, fig. 12) observations that there are regular relations between the major vowel parameters and the values of (at least) the first three formants it was decided that an interpolation was a good way to model the synthesis of the formant values from the vowel parameters. For the dimensions where there were three data points available (tongue height and tongue position) a quadratic interpolation was used, while for the dimensions where there were only two data points, a linear interpolation was used. The resulting equations are given in Table 1.

When an agent produces a formant pattern during the original imitation game, noise is added to it in order to make the production more realistic. When humans produce two instances of the same vowel, they will never produce exactly the same signal twice, even if one does not take the influences of context, speaker age, sex or dialect. Therefore the positions of the formants are shifted a random amount. For every formant a frequency shift is done in the following way:

$$\cancel{F}_i = F_i\ (1 + v_i\ )$$

where F_i is the frequency of the i^{th} formant as calculated by the synthesis function, and v_i is the amount of noise added to this formant. $\cancel{F}_i$ is the frequency of the formant after adding noise. The amount of noise is taken from the uniform distribution in the range

$$-\frac{\psi_{\mathrm{ac}}}{2} \leq v_i < \frac{\psi_{\mathrm{ac}}}{2},$$

where ι_{ac} is the maximum amount of noise that can be added or subtracted, which is a parameter of the system.

3.2 The perception model

The perception of the agents is based on the way humans perceive the formant patterns of vowels. It has been observed that human subjects perceive signals that have been generated with four or more formant peaks as similar to signals that have been generated with only two formant peaks. In these cases the position of the first formant peak is equal in both signals, while the position of the second peak in the simplified signal falls somewhere in between the second, third and fourth formant of the more complex signal (Carlson et al., 1970). The single peak that models the position of the higher formants is usually referred to as the effective second formant and written as F_2'. This effect is explained by the fact that the bandwidth of the detectors for higher frequencies seems to be higher than of the detectors for lower frequencies, such that detailed features at higher frequencies are blurred and for example multiple peaks can merge into one. It should be noted, however that not all vowels can be represented accurately by two formants. Especially high front vowels cause problems in this respect.

The frequency of the second formant is calculated by an algorithm that is a modified form of an algorithm devised by Mantakas et al. (1986) and that has also been described in Vallée (1994). This algorithm works on frequencies measured in the perceptually motivated Bark frequency scale. An equal distance in Barks implies a distance that is perceived as equal. In the models described in this paper, frequency expressed in Hertz is converted in frequency expressed in Bark by the following formula:

$$Bark = \begin{cases} \dfrac{\ln\left(\dfrac{Hertz}{271.32}\right)}{0.1719} + 2 & Hertz > 271.32 \\[2ex] \dfrac{Hertz - 51}{110} & Hertz \leq 271.32 \end{cases}$$

which is an approximation of the tables for Hertz to Bark conversion.

The algorithm is based on the notion of a *critical distance.* Following Vallée, this critical distance has been set to 3.5 Bark. Whenever the distance between F_2 and F_3 is higher than the critical distance, F_2' is taken to be equal to F_2. If the distance between F_2 and F_3 is larger than the critical distance, but the distance between F_2 and F_4 is larger than the critical distance, F_2' is taken to be a weighted sum of F_2 and F_3. When the distance between F_2 and F_4 is also smaller than the critical distance, then F_2' is taken to be the weighted sum of F_2 and F_3 if F_2 and F_3 are closer together than F_3 and F_4, otherwise F_2' is taken to be the weighted sum of F_3 and F_4. In a formula:

$$1. \quad F_2' = \begin{cases} F_2\,, & \text{if } F_3 - F_2 > c \\[1ex] \dfrac{(2-w_1)F_2 + w_1 F_3}{2}, & \text{if } F_3 - F_2 \leq c \text{ and } F_4 \text{ - } F_2 > c \\[2ex] \dfrac{w_2 F_2 + (2-w_2)F_3}{2} \text{-1}, & \text{if } F_4 - F_2 \leq c \text{ and } F_3 - F_2 < F_4 - F_3 \\[2ex] \dfrac{(2+w_2)F_3 - w_2 F_4}{2} \text{-1}, & \text{if } F_4 - F_2 \leq c \text{ and } F_3 - F_2 \geq F_4 - F_3 \end{cases}$$

The complicating factor in these formulas are the weight factors w_1 and w_2. In the original algorithm these were based on the prominence of the different

formants. As these are not calculated by the synthesis function, an estimate was made. A rough approximation is that the prominence of formant peaks increases whenever they are closer together. The two weights have therefore been estimated by:

$$w_1 = \frac{c - (F_3 - F_2)}{c}$$

$$w_2 = \frac{(F_4 - F_3) - (F_3 - F_2)}{F_4 - F_2}$$

Finally in the last two weighted sums in equation 1, a term −1 was added to the calculation of F_2'. This was done because otherwise a discontinuity would open between the frontmost front vowels and the rest of the vowel space. Although this corresponds to a real perceptual phenomenon (Schwartz, personal communication) it was found to interfere with the hill-climbing algorithm that was used to reconstruct the articulatory position of a vowel from a perceived signal.

The distance between two signals can be calculated in the F_1–F_2' space. However, it has been observed that changes in F_1 cause a greater difference in perception than changes in F_2'. Therefore a weighted Euclidean distance is used rather than a straight Euclidean distance:

$$D = \sqrt{(F_1^a - F_1^b)^2 + \lambda\left(F_2^{a\prime} - F_2^{b\prime}\right)^2}\,.$$

It has been found that $\lambda = 0.3$ results in the most realistic model of perception of vowel systems (Vallée, 1994, Schwartz et al., 1997b). Also there is independent evidence (Lindblom & Lubker, 1985) from the way displacements of the tongue are perceived that seem to indicate that distinctions in tongue height (which correspond roughly to distinctions in F_1) are perceived three times as accurately as distinctions in tongue position (which correspond to distinctions in F_2). This also supports a value of $\lambda \approx 0.3$.

Vowels are stored in terms of prototypes. The term prototypes is used here in the machine learning sense as the centre of a category. A new stimulus is classified as belonging to the category of the nearest prototype. This is a computationally efficient way of categorising. It is likely that prototypes play an important role in language and cognition (see e.g. Lakoff, 1987; Frieda et al., 1999). Using prototypes for learning and representation is therefore both a convenient as well as a realistic solution.

It must be stressed that the agents are language-independent. Both their articulation and perception are based on general properties of the human vocal tract and not on specific languages. They start out with an empty repertoire and learn vowels by playing imitation games with other agents without being biased towards the vowel systems of specific languages. Therefore conclusions about human languages in general can be drawn from these experiments, rather than only about some languages in particular.

3.3 The original imitation game

The interactions between the agents in the original imitation game have been described in great detail in (de Boer 2000). They will therefore only be described in a global way here. For each imitation game, two agents are taken from the population. One is assigned the role of initiator, the other is assigned the role of imitator. The initiator picks a random vowel from its repertoire, and synthesises this, while adding noise. This signal is then analysed by the imitator, who compares it to the acoustic prototypes of all the vowels in its repertoire, and picks the one that is closest to the perceived signal. It then synthesises this vowel, while also adding noise. The initiator in turn listens to this signal, compares it with all its vowel prototypes and picks its closest one. It then checks whether the vowel it recognised is the same as the one it originally produced. If this is the case, the imitation game is considered successful, otherwise it is a failure. This information is communicated to the imitator in a non-linguistic way. Note that the success of the imitation game depends on a number of factors: the shape of an individual's vowel system determines how easily vowels will be confused because of the noise that is added. But the success is also determined by how well the agent's vowel system conforms to the vowel system that is emerging in the population.

The use of this non-linguistic feedback might seem extremely unrealistic, as children do not receive such non-linguistic feedback every time they produce a sound. However, in order to learn which sound distinctions are meaningful in its language, a child has to get some feedback that links the sounds to objects and events in the world. Of course this feedback is not necessarily directly recognisable feedback. However, facial expressions or failure to achieve a communicative goal might as well serve to provide feedback. In any case, the functioning of the imitation game depends crucially on this feedback, but it is quite probable that children can do with much less intensive feedback.

The way the agents react to the imitation game determines how they can

learn the vowel system that is used by the other agents. Both the imitator and the initiator keep track of two counts for each of their vowels: the number of times it was used and the number of times it was successfully used. The ratio between these two values is a measure of how well the vowel fits into the vowel system that is emerging in the population. The imitator is the only agent that makes direct updates to its vowel system after the imitation game. If the imitation game was successful, it shifts the vowel it used a little bit closer to the signal it perceived. It does this by trying out small shifts in the positive and the negative direction of all the three articulatory parameters of this vowel. It listens to itself and the shift that brings its vowel closest to the signal it perceived is kept. The size of the shifts is a parameter of the system, and is set to 0.03 in all experiments described here.

If the imitation game was a failure, this can have two possible reasons. Either the vowel the imitator used is in the wrong place, and is not shared by the initiator, or the initiator has two vowels where the imitator has only one. These two cases can often be distinguished by the vowel's success/use ratio. Whenever this is low, this means that the imitation often fails, and that the vowel is probably just not very good. However, when the ratio is high, this means that the vowel must be a good imitation in a large number of cases, so that the most likely cause of failure of the game is a confusion with two nearby vowels in the initiator. Therefore, in the case the success/use ratio is high, (higher than 0.5 in fact) the imitator will add a new vowel that is a close imitation of the signal it heard. This is done by adding a central vowel (with all articulatory parameters set to 0.5) and repeatedly shifting it closer to the perceived signal, as described above. If the use/success ratio is low, on the other hand, the vowel that was used is just shifted closer to the perceived signal.

Other actions are undertaken every once in a while in order to clean up the agents' vowel systems. Vowels whose use/success ratio has fallen below a certain threshold (usually taken to be 0.7) after they have been used five times are considered hopeless and thrown out of the agents' vowel systems. Also, vowels that come so close together that they will too often be confused by acoustic noise (that is when the distance between their acoustic prototypes falls below the maximum shift that can be induced by adding noise) are merged. This is done in order to prevent clusters of bad vowels to emerge around the place where there should be only one. The acoustic and articulatory prototype of the new vowel are equal to those of the vowel with the highest use/success ratio and the other is thrown away. The use and success counts of the new vowel are the sums of the use and success counts of the two original vowels.

So far, nothing has been said about the way the imitation games are started and how the pressure to increase the size of the repertoires is implemented. In fact, the agents start out with an empty repertoire. In order for an initiator to start an imitation game, it has to add a random vowel if its repertoire is empty. Whenever the imitator's repertoire is empty, it tries to make a close imitation to the signal it heard by inserting a close imitation in the way that was described above. The pressure to extend the vowel repertoires is implemented by adding new random vowels to the agents' repertoires with a low probability (1% per imitation game).

3.4 The simplified imitation game

The simplified imitation game has been derived from the original imitation game in order to make it more analysable and in order to remove a number of assumptions about what the agents should be able to do. However, in order to do so the number of vowels, that is variable in the original imitation game, had to be fixed.

The architecture of the agents in this game is exactly the same as in the original imitation game. Their articulatory model, their perceptual model and the way they store vowels have remained the same. However, the agents are initialised with a repertoire of a predetermined number of vowels whose articulatory parameter values are chosen from the normal distribution with average of 0.5 and a standard deviation of 0.1. As the normal distribution goes from $-\infty$ to ∞, values that fall outside the range of 0 to 1 of legal articulatory parameter values are set to the nearest bound.

The interaction between two agents remains the same as in the original imitation game, only the way agents prepare for and react to the game are different, as well as the way agents are picked for the game. In fact, for this game only an initiator is picked. It then chooses a random vowel from its repertoire and adds random values to the articulatory parameters of this vowel, taken from the normal distribution with mean 0 and standard deviation 0.1 (again compensating for values that fall outside the legal range). It then plays imitation games with all other agents from the population and keeps track of how many games were successful. It then compares this value with the success value of the vowel it was using, which is now no longer a simple count, but a running average of the number of successful games per time it was chosen by an initiator. If the present success count is larger than the success count of the vowel, the articulatory parameters are updated as follows:

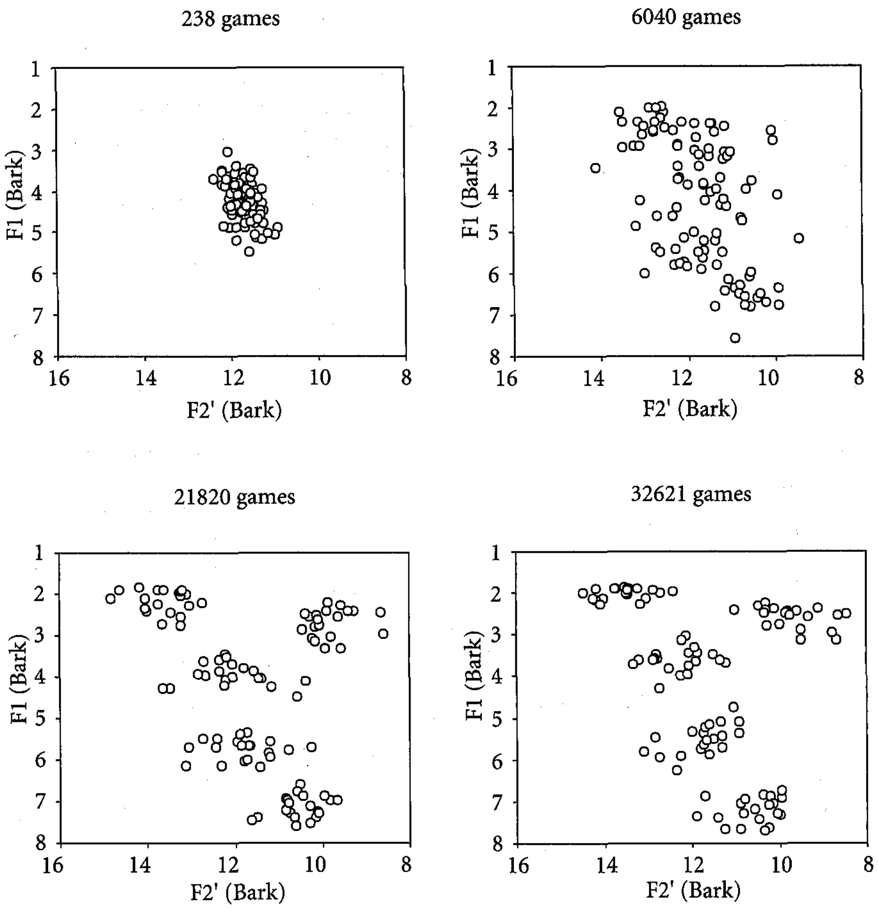

Figure 3. Emergence of a system of five vowels in the simplified imitation game

$$x_{new} = (1-\alpha)x_{original} + \alpha x_{used}$$

where x is any of the three articulatory parameters, $x_{original}$ is the original value, x_{used} is the value that is actually used in the imitation games, x_{new} is the new value and α is a weighting factor, set to 0.5 in the experiments that will be described.

Independent of the outcome of the imitation game, the success of the vowel is updated as follows:

$$success_{new} = (1-\alpha)success_{original} + \alpha \cdot success_{new}$$

where *success* indicates the success value and the subscripts are used as above. As will be shown below, this very simple algorithm also results in realistic vowel systems.

4. Results

The first result that is shown is the emergence of a five-vowel system in the simplified imitation game (Figure 3). In this figure, as in all subsequent pictures of vowel systems of populations of agents, all the acoustical prototypes of all agents are shown, plotted in the F_1-F_2' space with units in Barks. The simulation was run with a population of 20 agents and acoustic noise of 10%. This is a convenient 2-dimensional way of representing vowel systems in a population, because one can immediately see the shape as well as the clustering of the population's vowel system. It is also a reasonably realistic representation, as equal perceptual distances are represented as equal distances in the graph, if one allows for the fact that the horizontal dimension should actually be reduced to 30% of the length of the vertical dimension. As imitations are evaluated in the F_1-F_2' space, it is also a good representation of quality of imitation. It should be taken into account, however that due to articulatory limitations, the available acoustic space is roughly triangular, with the apex at the bottom of the frames.

The frames shown in Figure 3 show a typical evolution of a five-vowel system. One can see that initially (after 238 games) the vowel prototypes of the agents are distributed normally around the centre of the acoustic space. No structure can be discerned. After 6040 games, it can be observed that the agents' vowel prototypes have become dispersed about the available acoustic space, but a clear structure can still not be found. In the frame taken after 21 820 games however, five clear clusters are present, in which every agent in the population has a vowel prototype. This situation remains stable after this, although the positions of the clusters continue to shift and the clusters themselves become slightly more compact, as shown in the frame taken after 32 621 games.

The vowel system that eventually emerges does not look implausible, but does not occur in the $UPSID_{451}$ database. However, in other runs of the simulation symmetrical five vowel systems would emerge, which do occur very frequently in human languages. However, no data has been collected yet as to the frequencies of the different five vowel systems that emerge.

The results show that this very simple algorithm that uses only local interactions and no explicit optimisation is capable of generating dispersed and often realistic vowel systems. However, the disadvantage is that, like in Liljencrants and Lindblom's (1972) experiment the number of vowels needs to be determined beforehand. This is an unfortunate and unrealistic necessity, and the main focus in this paper will therefore be on the original imitation game, where the number of vowel prototypes per agent is also emergent. The simplified imitation game has been shown because it illustrates that emergent behaviour and self-organisation already occur in very simple systems, where almost no assumptions have been made about the cognitive abilities of the individual agents.

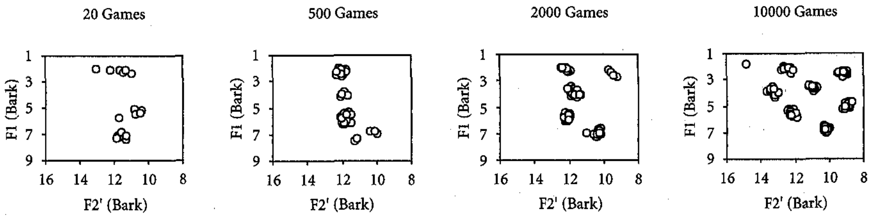

Figure 4. The emergence of a vowel system in the original imitation game.

The emergence of a vowel system in the original imitation game is illustrated in Figure 4. Here the emergence of a vowel system in a population of twenty agents under 10% of acoustic noise is shown. Two differences can be observed with figure three: the emergence is faster, and clusters appear directly instead of only at the end. Initially, all agents are empty, and in the first couple of imitation games, vowels are either added randomly to the agent's repertoire, or as the result of closely approximating another agent's vowel. Hence the small number of vowel prototypes in the leftmost frame of Figure 4 and the random dispersion of a number of small clusters. At the core of these clusters is a randomly added vowel, and the other prototypes around it are the other agent's approximations. However, as each agent has only one vowel prototype, no confusion is possible, and after every agent has added one vowel to its repertoire, no more adding is done. As all games are now successful, the only action agents undertake is shifting their prototypes closer together. However, new vowel prototypes are added at random every once in a while. Whenever the other agents can imitate these vowels, they spread through the population quickly. Hence after 500 games a small number of compact clusters has emerged. However, the vowel system at this stage is not yet very realistic. The distribution of the clusters is still rather haphazard. Only after 2000 games starts the available acoustic space to become filled with a number of dispersed clusters. Finally, after 10 000 games, a symmetrical 8-vowel system has emerged. Still, the vowel system is not completely evolved, yet. For example, not all agents share the high front vowel. Apparently it has been added recently and still has to spread through the population. Even when the whole vowel space has become filled with vowel prototypes, the vowel systems are not completely static. Vowel clusters can move slightly, or sometimes even merge or split. The vowel systems stay dispersed, and therefore continue to look realistic, though.

Is the realism of the emerged vowel systems only impressionistic, though, or do the regularities in the merged systems really correspond to the regularities found in human vowel systems? In order to assess this, the types of emerged systems and the frequencies of these types should be compared with types and frequencies of occurrence of human vowel systems. As has been mentioned above, Crothers (1978) derived a number of universals of human vowel systems, although they are rather universal *tendencies*. Not all human languages conform to all universals. It can be checked whether the emerged vowel systems also conform to these tendencies. Moreover, Schwartz et al. (1997a) have made a study of the frequencies of occurrence of different types of vowel systems, with which the frequency of occurrence of the emerged vowel systems can be compared.

In comparing different types of vowel systems, one does not look at the exact phonetic realisations of the vowels making up the system, but rather at the number of vowels, and the way these vowels are located in the available acoustic and articulatory space. Hence, the two three vowel systems /i/, /a/, /u/ and /e/, /a/, /o/ would both be classified as triangular three vowel systems and considered to be of the same type. On the other hand, the system consisting of /i/, /əd/ and /a/ is a vertical system and therefore considered to be different. It must be said that in deriving his universals, Crothers uses a slightly rougher classification than Schwartz et al. As the clusters that emerge in the simulations are still quite large, a rather rough classification has been adopted here as well.

It has been found (de Boer, 1999) by running many runs of the simulation for many different values of the acoustic noise parameter that most of the emerged vowel systems conform to most of Crothers' universals. Furthermore, there is a good fit between the frequencies with which the different types of vowel system emerge and the frequency with which they occur in human languages. The example for five vowel systems is illustrated in Figure 5. The data represented in this figure has been generated by running the original imitation game a hundred times for 25 000 games with acoustic noise set to 15% and a population of twenty agents. From these hundred runs, 49 ended with a vowel system with five vowels per agent. From these populations, a random agent was selected, this agent's vowel system was taken to be representative of the population and classified. It was found that 88% of the emerged systems consisted of the symmetrical five vowel system, 8% consisted of a system with two front vowels and one central vowel and 4% consisted of a system with two back vowels and a central vowel. This compares very favourably with the 89%, 5% and 5% (and 1% others) respectively found by Schwartz et al. (1997a). The first and second frame of Figure 5 (96%) also conform to Crothers' hierarchy (Figure 1). As the vowel systems of the rightmost frame also occur in human language, this also illustrates that not all human languages conform to Crothers' universals.

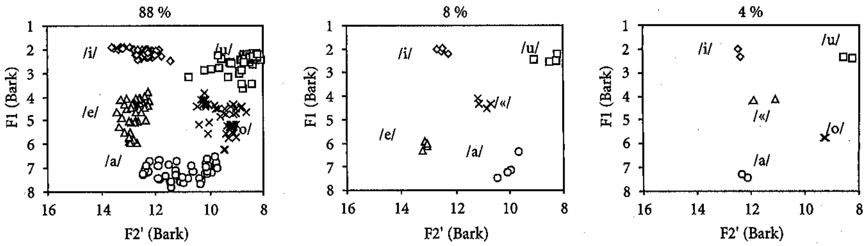

Figure 5. The classification of emerged five-vowel systems.

Admittedly, not all vowel system sizes have such good correspondence between the emerged and the human classifications, especially for three-vowel systems and systems with more than seven vowels, but a good match existed for all vowel system sizes between three and nine vowels. Furthermore, almost all emerged vowel systems conformed to the great majority of Crothers' universals of human vowel systems.

5. Discussion and Conclusion

These results show that realistic vowel systems can emerge in a population of agents as the result of self-organisation. The positions of the clusters are determined by the articulatory and perceptual constraints under which the agents operate. Although the emerged systems can be described perfectly in terms of distinctive features (Figure 6) these are not necessary for the emergence of the sound systems. As the match between the emerged systems and human vowel systems is so good, it is likely that the universals of human vowel

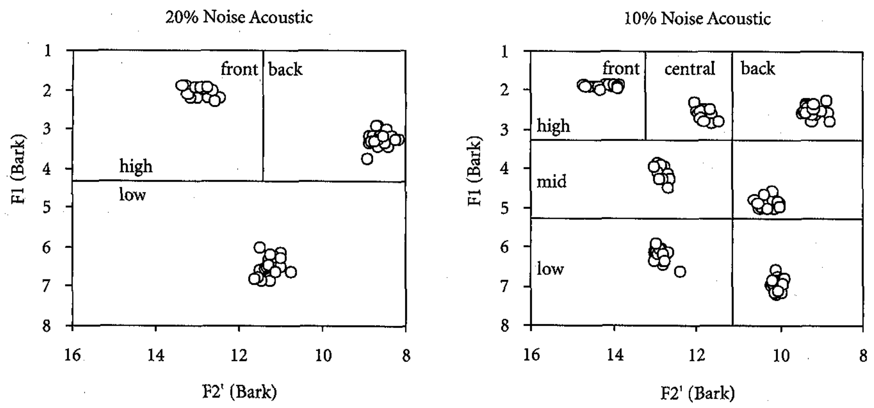

Figure 6. Emerged system described in terms of distinctive features.

systems are also the result of self-organisation, rather than the result of innate features and their markedness. This makes it unnecessary to find an evolutionary explanation for innate features of vowel systems.

Of course, the simulations are far from complete and realistic. The way a vowel system emerges in the simplified imitation game is more like the way vowel systems develop in infants (Kuhl & Meltzoff, 1996) than in the complex imitation game. In infants it seems that initially, they are not able to make distinctions between the different vowels, while after some exposure to language input, their range of vowel sounds expands in a way that is reminiscent of the way the agents' vowel repertoires expand in the simplified imitation game. Unfortunately in this simplified imitation game, the number of vowels has to be fixed beforehand. In the original imitation game, the number of vowels does not have to be fixed, but the development of vowel systems is not quite like either the way vowel systems develop in infants or the way vowel systems change historically. In future simulations, the learning properties of the simplified and original imitation games should be combined.

However, the fact that the emerged vowel systems are so much like the vowel systems that are found in human languages indicates that self-organisation is a very powerful mechanism that operates independently from the exact way in which the agents learn and interact. This is an extra indication that self-organisation probably also plays an important role in (the evolution of) human language.

Probably self-organisation also plays a role in the universals of more complex utterances. Its role has already been explored by Lindblom et al. (1984) in explaining phonemic coding in simple consonant-vowel syllables. Preliminary experiments have indicated that a version of the simplified imitation game is also able to make successful imitation emerge in populations of agents that imitate each other's complex utterances. The success of self-organisation in predicting the universals of human vowel systems also makes it worthwhile to explore the role of self-organisation in other parts of language, most notably in syntax. But at the moment anything said about self-organisation in syntax is pure speculation. Perhaps simplified computer models could help the investigation here as well.

As a conclusion it can be said that the universals of human vowel systems have emerged from the simulations presented here as the result of self-organisation under human-like constraints of perception and production, independent of the exact details of the interactions. Vowel systems that show the same universals as human vowel systems seem to be "attractors" of successful imitation in such systems. This observation makes it likely that these universals can be explained as the result of functional factors, rather than as the result of innate, biologically evolved mechanisms.

Notes

* This research was done at the artificial intelligence laboratory of the Vrije Universiteit Brussel. It is part of an ongoing research project into the origins of language and intelligence. Funding was provided by the GOA 2 project of the Vrije Universiteit Brussel. Part of the work was done at the Sony computer science laboratory in Paris, France.

References

Berrah, A.R. (1998) *Évolution Artificielle d'une Société d'Agents de Parole: Un Modèle pour l'Émergence du Code Phonétique*, Thèse de l'Institut National Polytechnique de Grenoble, Spécialité Sciences Cognitives.

Carlson, R., Granström, B. & Fant, G. (1970) Some studies concerning perception of isolated vowels, *Speech Transmission Laboratory-Quarterly Progress and Status Report (STL-QPSR)*, 2–3, pp. 19–35

Chomsky, N. & Halle, M. (1968) *The sound pattern of English*, MIT Press, Cambridge, Mass.

Crothers, J. (1978) Typology and Universals of Vowel systems. In J.H. Greenberg, C.A. Ferguson & E.A. Moravcsik (eds.) *Universals of Human Language*, Volume 2 Phonology, Stanford: Stanford University Press pp. 93–152.

de Boer, B.G. (1999) *Self Organisation in Vowel Systems*, PhD Thesis, AI-Lab, Vrije Universiteit Brussel.

de Boer, B. G. (2000) Emergence of vowel systems through self-organisation, *AI Communications* 13 27–39

Erwin, E. Obermayer, K. & Schulten, K. (1995) Models of Orientation and Oular Dominance Columns in the Visual Cortex: A Critical Comparison. Neural Computation 7, pp. 425–468.

Frieda, E. M., Walley, A. C., Flege J. E. & Sloane, M. E. (1999) Adults' perception of native and nonnative vowels: Implications for the perceptual magnet effect. Perception & Psychophysics 61(3), pp. 561–577

Glotin, H. (1995) *La Vie Artificielle d'une société de robots parlants: émergence et changement du code phonétique.* DEA sciences cognitives-Institut National Polytechnique de Grenoble.

Glotin, H. & Laboissière R. (1996) Emergence du code phonétique dans une societe de robots parlants. *Actes de la Conférence de Rochebrune 1996 : du Collectif au social,* Ecole Nationale Supérieure des Télécommunications — Paris.

Hurford, J. R, Studdert-Kennedy, M. & Knight, C. (eds.) (1998) *Approaches to the Evolution of Language* (selected papers from the 2nd International Conference on the evolution of Language, London, April 6–9 1998), Cambridge: Cambridge University Press

Jakobson, R. & Halle, M. (1956) *Fundamentals of Language,* the Hague: Mouton & Co.

Kuhl, P. K. & Meltzoff A. N. (1996) Infant vocalization in response to speech: Vocal imitation and developmental change, The journal of the Acoustical Society of America 100 (4) pp.2425–2438.

Ladefoged, Peter (1981), *Preliminaries to Linguistic Phonetics,* Midway Reprint, The University of Chicago Press.

Ladefoged, P. & Maddieson, I. (1996) *The Sounds of the World's Languages,* Oxford: Blackwell.

Lakoff, G. (1987) *Women, fire, and dangerous things: what categories reveal about the mind.* Chicago: Chicago University Press.

Liljencrants, L. & Lindblom, B.(1972) Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language* 48 pp. 839–862.

Lindblom, B. & Lubker J. (1985) The Speech Homunculus and a Problem of Phonetic Linguistics. In V. A. Fromkin (ed.) *Phonetic Linguistics: essays in honor of Peter Ladefoged,* Orlando: Academic Press pp. 169–192.

Lindblom, B., MacNeilage, P. & Studdert-Kennedy, M. (1984) Self-organizing processes and the explanation of language universals. In B. Butterworth, B. Comrie & Ö. Dahl (eds.) *Explanations for language universals,* Walter de Gruyter & Co. pp. 181–203.

Maddieson, Ian (1984) *Patterns of sounds,* Cambridge University Press.

Maddieson, Ian & Kristin Precoda (1990) Updating UPSID. In *UCLA Working Papers in Phonetics* 74, pp. 104–111.

Mantakas, M, Schwartz, J. L. & Escudier, P. (1986) *Modèle de prédiction du 'deuxiéme formant effectif' F_2' — application à l'étude de la labialité des voyelles avant du français.* In Proceedings of the 15th journées d'étude sur la parole. Société Française d'Acoustique, pp. 157–161.

Nicolis, G. & Prigogine, I. (1977) *Self-organization in non-equilibrium systems,* New York: John Wiley.

Petitot-Cocorda, J. *Les catastrophes de la parole; de Roman Jakobson à René Thom,* Paris: Maloine

Schwartz, J. L., Boë, L. J, Vallée N. & Abry, C. (1997a), Major trends in vowel system inventories. *Journal of Phonetics* 25, pp. 233–253

Schwartz, J. L., Boë, L. J., Vallée N. & Abry, C. (1997b), The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics* 25, pp. 255–286.

Steels, Luc (1996) The Spontaneous Self-organization of an Adaptive Language. In S. Muggleton (ed.) *Machine Intelligence* 15.

Steels, L. (1997) The Synthetic Modelling of Language Origins, *Evolution of Communication* 1(1): pp. 1–34.

Steels, L. (1998a) The Origins of Ontologies and Communication Conventions in Multi-Agent Systems. *Autonomous Agents and Multi-Agent Systems* 1, 169–194

Steels, Luc (1998b) The origins of syntax in visually grounded robotic agents. *Artificial Intelligence* 103(1–2) pp. 133–156.

Vallée, N. (1994) *Systèmes vocaliques: de la typologie aux prédictions,* Thèse préparée au sein de l'Institut de la Communication Parlée (Grenoble-URA C. N. R. S. no 368).

Vogt, P. (1998) Perceptual grounding in robots. In: Birk, A. & J. Demiris (eds.) *Proceedings of the 6th European Workshop on Learning Robots 1997. Lecture Notes on Artificial Intelligence* 1545. Berlin: Springer-Verlag.

Wildgen, W. Basic Principles of Self-Organization in Language, In: Haken, H. & Stadtler, M. (eds.) *Synergetics of Cognition,* Berlin: Springer Verlag, pp. 415–426

Wittgenstein, Ludwig (1967) *Philosophische Untersuchungen,* Frankfurt: Suhrkamp.

Author's address

Bart de Boer
AI-Lab
Vrije Universiteit Brussel
Pleinlaan 2
1050 Brussel
Belgium
bartb@arti.vub.ac.be