Analogies between Genome and Language Evolution

Luc Steels[1,2]
[1] Sony CSL - Paris - 6 Rue Amyot, 75005 Paris
[2] University of Brussels (VUB AI Lab)
E-mail: steels@arti.vub.ac.be

Abstract

The paper develops an analogy between genomic evolution and language evolution, as it has been observed in the historical change of languages through time. The analogy suggests a reconceptualisation of evolution as a process that makes implicit meanings or functions explicit.

Keywords: Language evolution. Evolution of communication. Cultural evolution and learning.

Introduction

In November 1998, a small workshop involving linguists, biologists, and Artificial life researchers was held in Paris with the goal of exploring analogies between language evolution and genomic evolution.[1] As one might expect, the discussion was both enormously stimulating but also very inconclusive. It brought out great gaps between the fields, partly caused by a lack of clear theories, particularly for language evolution. Analogies are very risky. Nevertheless, they play an important role in scientific discovery (such as the analogy between planets circling around the sun and electrons orbiting the nucleus). In the best case, they lead to a conceptual revision, both of the source of the analogy and its target. The analogy between language evolution and species evolution was first proposed by Darwin, who was strongly influenced by August Schleicher, Ernst Haeckel, and other 19th century linguists who viewed language as a living system (Richards, 1987). More recently, the syntactic structure of genomes is being described using the same formalisms as used in linguistics, and genomic evolution is being modeled in terms of changes to formal grammars (see e.g. (Dassow, 1996)). With the background of better Artificial life models of language evolution (as surveyed for example in (Cangelosi and Parisi, 2001) and (Steels, 2003)) and better knowledge of the functions of the genome in development and genomic evolution, I develop in this paper another kind of analogy between language systems and genomes, emphasising meaning or function.

The paper is intended to be a discussion paper at a conceptual level. The main purpose is to formulate constraints and issues that must be addressed in models of language or genomic evolution. At the same time, the paper provides background and justification for computational and robotic experiments discussed in other more technical papers ((Steels and Kaplan, 2000), (Steels, 2003), (Steels, 2004)). There are two major points: First I emphasise the benefit of looking at the whole system (form, meaning, and effect in the case of language; genes, biochemical function, and structure/behavior in the case of genomes), as opposed to only focusing on the evolution of syntax. Second I will emphasize that both language evolution and genomic evolution are concerned with making certain meanings/functions explicit which were implicit before, or vice-versa. The big issue is then: how we can understand the mechanisms underlying this process and how we can synthesise them in artificial systems.

Evolution of language and languages

A distinction must first of all be made between the evolution of language, in the sense of the origins of language, and the evolution of languages throughout human history. This is analogous to the distinction be-

[1] Participants included F. Cambien, P. Hogeweg, T. Ikegami, D. Krakauer, T. Kuteva, L. Steels, E. Száthmary, B. Vittori, and G. Weisbuch

tween the origins of life itself and the subsequent evolution of living organisms over millions of years. Investigations into the evolution of language focus on finding developmental histories of how different brain areas could have become recruited for language and what factors might have caused verbal behavior to become such an important part of human activity. The work of Deacon (Deacon, 1997) is representative for this research challenge.

Investigations into the evolution of languages take the form of empirical investigations surveying the actual change in language, for example from Latin to French, Spanish, or Italian (see e.g. (Hopper and Traugott, 2003), (Heine and Hnnemeyer, 1991)), and theoretical investigations trying to identify and/or simulate the cognitive processes that give rise to these changes (see e.g. (Heine, 1997), (Steels, 2004)). It is generally assumed that language change cannot be based on genetic evolution because (1) it is very fast compared to genetic evolution, and (2) a person born in one linguistic community can learn the language of another community quite easily, even though the earlier one starts the better.

There are possibly very strong relations between the original evolution of language and the subsequent evolution of languages, in line with the uniformitarian hypothesis (adopted by Lyell in geology and Darwin in biology): The same processes that have molded languages throughout history must also have been playing a role in the genesis of language *ab initio*, and indeed they have been observed when a lexical language (pidgin) evolves into a creole (DeGraff, 1999). There also appears to be obvious connections between language learning and language change, in the sense that the cognitive operators which have been hypothesised as driving language change, are highly relevant to the ones underlying the socio-cultural learning of language (Heine, 1997). This paper focuses only on the evolution of languages without exploring these additional ramifications.

Defining Language Evolution

In order to characterise language evolution more precisely, I am going to take a functionalist point of view, which means that language is primarily seen as a vehicle for communication, and so its origins and evolution fit within the general process of evolving communication systems. Communication is here defined as the process whereby one agent (the speaker) deliberately influences the behavior of another agent (the listener) using (conventionalised) signs. Language therefore involves three aspects: forms, meanings and effects.

- The *forms* of language are sounds, words, word order patterns, intonation, stress, etc. They are the observable building blocks with which utterances are made.
- The *meanings* are what is expressed by utterances. Meanings are here defined as distinctions relevant to the agent-environment interaction. For example, the distinction between red, green, and orange traffic lights is relevant for deciding whether to cross the street or not. Meanings are assumed to be coded as information states so that they can play a role in semantic processes, instantiated as transformations over information states.
- The *effects* of an utterance are the behaviors carried out by the listener as a result of the meanings deduced from the form of an utterance. The most basic effect of language is to draw attention to an object or event in the world but many other effects are possible.

Consider a scene where two people are walking towards a bus stop. One looks behind and suddenly shouts "the bus", after which both start to run. The forms here are the words "the" and "bus" put in a particular order. The meanings include (1) a specific class of autonomously moving objects (buses), and (2) an indication (using the word "the") that there is a unique bus being expected in the present context. The effect of this utterance is to draw the attention of the listener to the fact that a bus is approaching and to take immediate action to catch it.

There are four important properties of human natural languages which are of crucial importance for the present discussion:

(1) Typically the form of the utterance only gives a hint about expected behavior. It influences behaviors which might already be going on anyway, without fully causing or determining them. In the example above, the speaker did not say that the listener should start running or whether a bus was approaching, she just said "the bus". The participants were already walking towards the bus and shared the context and goals. Natural language is therefore not a code in the sense of Shannon, which simply translates information from one form into another (Sperber and Wilson, 1987). Part of the meaning must be reconstructed based on the shared situation, common ground, joint attention, inference, etc.

This is why it is extraordinarily difficult for computers and robots to parse and interpret human language and it raises doubts whether information theory is a good framework for studying human natural language and language evolution.

(2) The relation between form, meaning, and effect is very indirect and multi-layered. Several words and grammatical constructions often collaborate in a non-modular way to constrain the possible meanings of the utterance, and there is a multi-layered hierarchical structure with certain words and constructions having a purely regulatory effect on the meanings and effects of others. For example, the word "back" has many meanings: a body part ("my back hurts"), a spatial area ("in the back of the car"), a temporal relation ("back in the good old days"), an adverbial particle ("I will be back"). The syntactic context and semantic expectations help the listener to pick out effortlessly the intended meaning. The influence by the meanings of an utterance on action selection (the 'illocutionary force') is even more determined by the context. For example, whether the utterance "the bus" evokes running or not depends entirely on what is happening in the present situation.

(3) There are important differences between languages in what meanings they make explicit, either in the lexicon or in the grammar. For example, European languages typically express tense and aspect (present/past/future, progressive, perfective/imperfective) through morphology and grammatical constructions. Compare: "I will write a letter" (future) with "I was writing a letter" (past imperfective). In Chinese, tense is not explicitly expressed grammatically but must be circumscribed indirectly, or inferred from the context, even though aspect is made explicit (for example with the particle -le for perfective aspect).

Thus the following sentence is unclear whether the washing was in the past or the present.

```
Akiu xi-zhe na jian dayi.
Akiu wash-prg that clothing-cl coat
Akiu is/was washing that coat.
```

This example illustrates also that Chinese, similar to most African languages, makes a distinction between different classes of nouns which are expressed through classifiers like "jian". English weakly uses a distinction between male/female/neuter, but otherwise does not express the distinctions implied by the Chinese classifiers at all.

(4) Finally, there is substantial evidence from all the world's languages and over all periods of recorded history, that profound changes take place, both in what meanings are made explicit and in how they are made explicit. For example, many languages (like Latin, Old-Germanic, Chinese, Polish) do not have a separate syntactic class of articles (like "the", "a", etc.) to express determination (definiteness/indefiniteness with respect to present context, quantification, etc.). It has been shown that a grammatical system for determiners may evolve in a language, as indeed it did in most languages that evolved from Latin (French, Italian) or from Old-Germanic (German, Dutch, Danish), typically by changing the form and function of demonstrative pronouns, like "that" => "the" or "ille" (Latin) => "le" (French).

The process by which new grammatical subsystems arise is generally known as grammaticalisation (Hopper and Traugott, 2003) and is discussed further below. It has also been shown that certain grammatical systems may disappear, at which point their function is totally lost or it is taken over by another system which develops often in competition with the first. A well known example is the case system in old English (consisting of morphological affixes or inflections to make the role of the referent of a noun phrase in an event explicit, as in German der/dem/den/des). The case system of early Old English was similar to that of Latin or (old-)German in complexity but largely disappeared by the advent of Middle English. This meant that other means had to be found in order to express these roles, leading to a tightening of word order and the extended use of prepositions. These grammaticalisation phenomena are precisely the processes that any theory of language evolution must explain.

In conclusion, we can view language evolution as follows:

> Language evolution is the process whereby meanings which were implicit, become explicit or vice-versa.

The Analogy with Genomic Evolution

An organism's genome and its role in the development and functioning of an organism is of course in many (if not most) respects very different from an utterance or set of utterances. Nevertheless we can view the genome as a kind of communication which influences

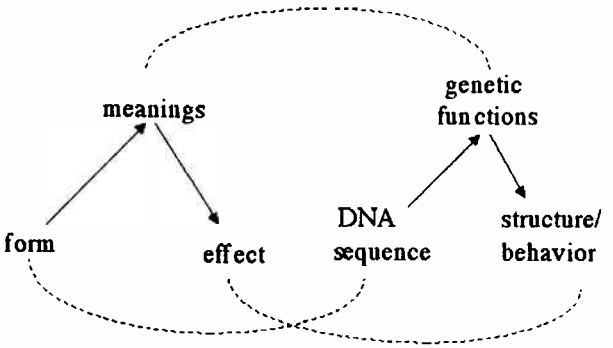

Figure 1: An analogy is suggested between the form/meaning/effect of language and the DNA/gene function/structure relation in development of organisms.

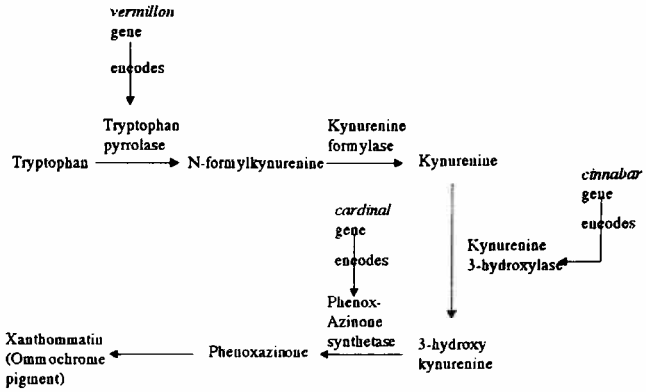

Figure 2: Example of a simple pathway for the synthesis of color pigments in the *Drosophilia* eye. Most biochemical pathways are much more complex, forming networks rather than linear chains.

how the organism is to develop, maintain itself, and behave. Just as in the case of language and in line with functional genomics, we will adopt a functionalist viewpoint, looking at the whole process from genomes to behaving organisms. The cores of this system are the biochemical pathways that determine the development of cells, tissues, and organs, and their structural maintenance and functioning. For example, the synthesis of ommochrome pigments in the *Drosophilia* eye is based on the 'tryptophan degradation pathway' schematically shown in figure 2 ((Wilkins, 2002) p. 104). Each step in a pathway synthesises molecular substrates, regulated by enzymes acting as catalysts. The function of genes is to act as such regulators, co-determining whether a transition takes place or not. Some of them have meta-functions, regulating the activities of other genes, or repairing gene copying. But other factors may intervene in the success of a transition as well, for example, certain substrates or catalysts might have to be provided by the environment or maybe byproducts of earlier biochemical transitions.

We can identify three aspects to genome-steered development, analogous to the form, meaning, and effect of language (see figure 1).

- An organism's genetic material, the DNA sequence, is similar to the form aspect of an utterance.
- The functions of genes in establishing transitions in biochemical pathways play the same role as the meanings of utterances.
- The effects of gene function (in specific contexts) are the behaviors and structures of the cells, tissues, and organs that allow the organism to function in a particular way.

Given this analogy, we can see that the genomic system has some of the same properties as natural language based communication systems: (1) Genes influence the transitions in biochemical pathways but are not necessarily the sole cause or controller of a transition. Substances available in the environment and even environmental stimuli processed by specific sensors, such as a pheromones, may co-determine whether a morphogenetic pathway unfolds. (2) The relation between the genome, the functions of the genes in orchestrating particular biochemical pathways, and the resulting structure and behavior of the organism is very indirect and multi-layered. For example, many genes are concerned with setting up the context for others, forming gene regulation networks with activators, inhibitors, pleotropic regulators, etc. Just as words and grammatical constructions are polysemous and ambiguous, the same gene product may be used multiple times in different cell or tissue types and at different times. (3) Differences between species (and even among members of the same species) concern what transitions in biochemical pathways are explicitly regulated or influenced by genes or not, in other words whether a certain biological function is genetically determined. In the most obvious case, this regulation establishes whether the biochemical pathway

can complete its course or not, and hence whether the organism has certain structures or behaviors. But it can also be a matter of variation in the probability or speed of certain transitions. (4) There are profound changes in whether certain biochemical transitions are mediated by gene-encoded enzymes or not. Ernst Mayr already pointed out that morphological variation can not only occur by genetic variation, typically in highly canalised systems, but that there is also non-genetic variation, for example due to environmentally induced plastic responses (Mayr, 1963).

Based on these observations, it makes sense to view genomic evolution as analogous to language evolution:

> Genomic evolution is the process whereby transitions in biochemical pathways which where were implicit, become (genetically) explicit, or vice-versa.

Empirical Data on Language Evolution

A lot is known from empirical observations by historical linguists how language evolution takes place ((DeGraff, 1999), (Heine and Hnnemeyer, 1991)), even though there are today hardly any good theoretical models of the causal mechanisms that underly them. Basically, the following five phenomena have been identified.

1. *Lexicalisation* Lexicalisation is the process by which a word becomes associated with a (new) meaning. The activities of human beings are in constant flux and new behaviors come up all the time, for example driven by the development of new technologies. Hence new distinctions become relevant and they may lead to a subsequent need to express them in verbal interactions. There may also be a desire to express existing meanings in new ways in order to 'keep the listener on her toes'. In principle, any word can be associated with any meaning because there is an arbitrary relation between form and meaning in language. But usually the word form is not constructed *de novo*, but rather an existing word whose meaning has some analogical or metaphorical relationship to the new meaning, is recruited. A listener must be able to guess the (new) meaning of a word and if the word has already a meaning which is close that becomes more likely to succeed. For example, the color word "orange" is adopted by metonymy from the orange fruit.

2. *Syntacticisation* New meanings (or existing meanings) are not necessarily expressed by new word forms. They may also be expressed by using suprasegmental form characteristics, such as the ordering of words, intonation, stress on certain words or syllables, etc. Moreover words do not occur in isolation but are linked to each other in syntactic contexts which determine part of the interpretation process. For example, in the utterance "the oldest girl sent a letter to her father" we know that "oldest" and "girl" cooperate to identify a particular referent, that "her" probably refers to the same referent, that "her" and "father" cooperate to refer to another person, etc. Syntactic structures imply that words or word groups become members of particular syntactic categories (such as adjectives, nouns, noun phrases, etc.) which constrain their combination and thus their interpretation.

3. *Grammaticalisation* Grammaticalisation is the process whereby a word or syntactic structure shifts to carry a grammatical function, in the sense that it looses some of its original lexical meaning (a process known as semantic bleaching) to express a more abstract meaning or gain a new one, and it looses some of its original syntactic properties (syntactic bleaching) (Hopper and Traugott, 2003). In the process of grammaticalisation, a new syntactic category may emerge or the recruited word or syntactic construction may become assigned to another syntactic category, which often implies a new syntactic context, e.g. different word order, ability to engage in morphological variation, etc.

For example, a verb of volition ("will") has become a future tense auxiliary in English (as in "It will rain tomorrow") (Bybee and W.Pagliuca, 1994). The verb "will" was originally a main verb with the meaning of "want" (it still is in Dutch), but it became recruited to express the more abstract sense 'future' and shifted syntactic category to auxiliary. This implied a specific position in the sentence (for example, before the subject in interrogative sentences, as in "Will it rain tomorrow?") and loss of the ability to have direct objects (one can no longer say "I will a book" in the sense of "I want a book").

4. *Cliticisation and Affixation* Once a lexical form has become grammatical it has a strong tendency to loose some of its original phonological structure and become first a clitic, that is a highly simplified word form which must occur next to its host (such as "ll" in "I'll see him tomorrow"). In a further process of phonetic erosion and simplification, the clitic may gradually become a morphological affix that is an inherent part of

the word, such as "-ed" in "walk-ed".

5. *Deletion* A final step is that the clitic or affix becomes so weak that it is no longer clearly audible and looses its function altogether. In that case the construction may progressively disappear and the cycle repeats itself, beginning with a new lexical item that is recruited to serve the same purpose. Often some debris of earlier evolutions is left behind and then later becomes available for recruitment to new functions.

These steps in language evolution (usually called grammaticalisation chains or 'clines') do not occur in any kind of predictable time frame and often there is competition between newer forms and existing ones. For example, for the expression of past tense in English, a system of form variation ("do" vs. "did") still co-exists with the use of a morphological affix ("walk" vs. "walked") and an auxiliary ("I wrote him" vs. "I did write him"). Nevertheless there is considerable regularity in evolutionary paths. For example, the recruitment of a verb of volition for expressing future is found in totally diverse language families (Bybee and W.Pagliuca, 1994) There is also a consensus that this type of evolution is basically uni-directional: from lexicalisation to grammaticalisation and cliticisation or affixation.

Historical linguists usually take a global view when tracking historical changes in language. But language does not exist as an abstract entity separate from the use of language in situated interactions between speakers and listeners. It is rather like a species, as Darwin already suggested. Each language user has his or her own private knowledge of the language, which may differ considerably from others depending on the history of interactions of the individual and the network structure of the population. It it the cumulative microchanges made by speakers and listeners that cause the global evolution in the language, which can usually only be observed in hindsight.

This raises two crucial questions (1) What are the social and physical behaviors and cognitive operations that individual language users carry out so that their net effect gives rise to the evolutionary phenomena observed and (2) what are the forces that drive the grammatical pathways forward?

The first question can only be adequately addressed against the background of theories of the cognitive processes of language understanding and production, for which we have today better and better (computational) models (Levelt, 1989). Next, we have to add to the normal operation of these processes the cognitive mechanisms that language users employ to perform 'language engineering': recruit existing words and syntactic structures for new meanings, to guess meanings of unknown elements, to push a lexical item towards a more grammatical item, etc. Heine (Heine, 1997) is one of the few diachronic linguists who have attempted to circumscribe these language molding operations but much work remains to operationalise them and embed the required mechanisms in a global theory of language use.

With respect to the second question, many historical linguists implicitly adopt the following hypothesis (see e.g. (Hopper and Traugott, 2003)): The goal of a communication between two participants (speaker and listener) is to reach communicative success with minimal effort. Making additional meanings explicit may help because it restricts the possible set of interpretations. The speaker also forces the listener to make similar distinctions in a particular situation and hence making meanings explicit helps to coordinate them in a population. On the other hand, there is a cost associated with expressing meaning, in terms of memory, processing, and learning. Hence languages try to find a balance between expressive power and effort. There is no unique solution for this multiple constraint satisfaction problem and choices are necessarily made, either based on historical accidents which continue to propagate or because certain meanings are more important to the culture in which the language developed.

Parallels with Genomic Evolution

The mechanisms underlying genomic evolution are progressively being unraveled thanks to the growing availability of complete genomes and techniques for tracking the function of individual genes in the development of cells and cellular structures. Earlier views based on the notion that genetic material consists of a set of individual, well-defined genes, each with a discrete function, have been abandoned in the face of the discovery of complex gene-regulatory networks. By implication, genetic evolution, which used to be thought of in terms of the cumulative stochastic operations over genes (mutation, duplication, recombination), is now viewed as a much richer process that can be understood in terms of the recruitment and co-optation of genes and gene regulatory networks for new functions (Carroll, 2001). In this sense, genomic evolution is based on 'genetic engineering' processes that are analogous to the

'language engineering' activities underlying language evolution. For example, Radman (Radman, 1999) and colleagues have shown that genetic mutations and recombinations are not purely stochastic events but may partly take place under genomic regulatory control, induced by environmental stress such as increased errors in DNA-copying (SOS response). The environment in other words triggers the need for exploring ways in which certain functions become explicitly encoded for. Similar to lexicalisation, an existing gene (already used in another pathway) may become inserted in a pathway and thus leads to a new function that then undergoes selection (Wilkins, 2002).

Conclusion

One of the goals of Artificial Life is to synthesise in artificial systems evolutionary phenomena such as the ones that are discussed in this paper. It is clear that this is going to be extraordinarily difficult in the case of language evolution (and even more difficult for 'realistic' genomic evolution). Verbal communication engages all areas of cognition and is situated within concrete settings experienced through embodied interactions. It is impossible to capture all that in computational models. The use of robots (as advocated in (Steels, 2001)) already introduces the real-world interaction and the embodiment, but brings of course additional complexities to set up and carry out experiments.

The analogy between language evolution and genomic evolution has a heuristic value for developing frameworks and artificial life experiments, both in the domain of language evolution and in the domain of genomic evolution. When we view evolution as 'making explicit meanings or functions which were implicit before' or on the contrary 'eliminating explicit expression when no longer needed', then interesting parallels and similar questions start to appear, such as: how may intermediary levels appear, how can basic structures (such as intermediary control genes in the genome or basic sentence patterns in language) be conserved, despite constant change, etc. We need to understand the 'genetic engineering' that organisms use to adapt themselves to environmental conditions, just as a theory of language evolution requires mapping out the 'language engineering' that language users engage in to adapt the language to their needs. Of course there are also tremendous differences between language and genome evolution but that should not prevent us from exploiting the analogies.

Acknowledgement

This reseach was conducted at the Sony Computer Science Laboratory. Additional funding came from the CNRS OHLL project, the ESF OMLL project, and the EU FET program on ECAgents. I thank C.L. Nehaniv for many interesting comments in his review of the paper.

References

Bybee, J. L. Perkins, R. and W.Pagliuca (1994). *The Evolution of Grammar. Tense, Aspect, and Modality in the Languages of the World.* Chicago, University of Chicago Press.

Cangelosi, A. and Parisi, D. (2001). *Simulating the Evolution of Language.* Berlin: Springer-Verlag.

Carroll, S., J. G. S. W. (2001). *From DNA to Diversity. Molecular Genetics and The Evolution of Animal Design.* London, Blackwell Science.

Dassow, J., V. M. (1996). Evolutionary grammars: A grammatical model for genome evolution. In Dassow, e., editor, *German Conference on Bioinformatics*, pages 199–209. Berlin, Springer-Verlag.

Deacon, T. (1997). *The Symbolic Species; The Coevolution of Language and Brain.* New York, W. Norton and Company.

DeGraff, M. (1999). *Language Creation and Language Change. Creolization, Diachrony, and Development.* Cambridge Ma, The MIT Press.

Heine, B. (1997). *Cognitive Foundations of Grammar.* Oxford. Oxford University Press.

Heine, B., U. C. and Hnnemeyer, F. (1991). *Grammaticalization: A Conceptual Framework.* Chicago. The University of Chicago Press.

Hopper, P. and Traugott, E. (2003). *Grammaticalization.* Cambridge.Cambridge University Press, Cambridge.

Levelt, W. (1989). *Speaking: From intention to articulation.* Cambridge.Cambridge University Press, Cambridge.

Mayr, E. (1963). *Animal Species and Evolution.* Cambridge Ma. Harvard University Press.

Radman, M. (1999). Mutation: enzymes of evolutionary change. *Nature*, 401:866–869.

Richards, R. (1987). *Darwin and the emergence of evolutionary theories of mind and behavior.* Chicago, The University of Chicago Press, Chicago.

Sperber, D. and Wilson, D. (1987). *Relevance, communication and cognition.* Oxford, Basil Blackwell.

Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent Systems*, September/October:16–22.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Science*, 7,7:308–312.

Steels, L. (2004). Self-organising grammars in embodied situated language games. In Hurford, e., editor, *Proceedings Evolution of Language V*, page 40. Leipzig, MPI.

Steels, L. and Kaplan, F. (2000). Collective learning and semiotic dynamics. In Foreano, D., N. J.-D. and Mondada, F., editors, *Advances in Artificial Life (ECAL 99)*, pages 679–688. Berlin, Springer-Verlag.

Wilkins, A. (2002). *The Evolution of Developmental Pathways.* Sunderland Ma, Sinauer Assoc.