Seminar für Computerlinguistik
Institut für allgemeine und angewandte Sprachwissenschaft
Ruprecht-Karls-Universität Heidelberg

Magisterarbeit

Computer Simulation of Language Evolution

Steffen Eger

26 January 2007

Supervisors:
Dr. Markus Demleitner
and
Prof. Dr. Peter Hellwig

Abstract

This master thesis presents a computational approach to the problem of language evolution, the study of how language emerges from a situation in which there is no language. The model employed is inspired by the work done in Livingstone (2002) and Livingstone and Fyfe (1999). The structure of the thesis is as follows: Chapter 2 introduces the basic terminology and results of dialectology and language change. Chapter 3 gives an overview of the simulation model and the details of implementation. Chapter 4 provides an analysis of the model's parameters and discusses the results observable for the 'basic' implementation. Chapter 5 describes the setup and outcomes of genuinely novel experiments; these include the study of the emergence of simple compositional grammars. Chapter 6 recapitulates the findings obtained and relates these to questions of general linguistic importance.

Preface

Ah, distinctly I remember it was in the bleak December,
And each separate dying ember wrought its ghost upon the floor.
(Edgar Allan Poe, *The Raven*).

Orthographic conventions

The text presented in this master thesis is supposed to be written following the conventions of American English. This means that the spelling, say, *favor* was favored over other possible spellings, such as *favour*. However, since the distinction between the varieties of English is, in the view of the author of this thesis, a matter of tendencies and probabilities rather than an issue of binary oppositions, certain liberties were exercised - possibly depending on the mood and disposal of the author - as to the spelling of such words as, say, *analyse* versus *analyze*. Concerning the use of pronouns, the feminine version - contrary to the business of most other present day literature, which endeavor to establish 'some sort of equality' - was employed when the confusion of the (current) reader was thereby not unnecessarily provoked. Exceptions in this sense were cases like *the learner updates the weights of his net* where *the learner* most likely - and actually - refers to an artificial agent but the term is general enough to not exclude the possibility of *the learner* being human. I felt that such instances would cause the alienation of the *current* reader unless the traditional, longer established masculine pronoun were chosen. References to an *agent* or *agents* were made - the idea of an 'agent' being that of a non-human, artificial machine in this work's context - in the neutral form because it was not believed that the employment of, for example, *she* was justified in these cases; even less so that of *he* and associates. Punctuation, in particular comma placement, was exerted with the proverb

"if some is good, more is better" in mind. Furthermore, it is a pity that the common English writer seems to do so well without the colon; I dearly missed it. Finally, the excessive usage of parentheses is a thorn in the eye and an obstacle to fluent reading (but, unfortunately, the author of this work seems to have the persistent belief that he has still something more to say).

Heidelberg, January 11, 2007

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Since language is so important for defining what a human being is, human beings have a natural tendency to propose theories and evolutionary scenarios for its origin and evolution (Cangelosi and Parisi, 2002: p.4).

1.1 Language evolution

"The study of language origins and evolution is the study of how language emerges from a situation in which there is no language" (Cangelosi and Parisi, 2002: p.3). In particular, questions are asked about the prerequisites enabling such a transition, and about its implementation. When thus delimiting the scope of this work's subject area, one needs to emphasize that language evolution is neither equivalent to historical linguistics nor to child language acquisition; both presuppose situations in which language already exists. However, these disciplines - alongside others, such as developmental psycholinguistics, paleoanthropology, evolutionary biology, anthropology, and the neurosciences - may have something to say about it.

While this rich interdisciplinary basis may render the study of language evolution colorful and interesting, it may also be a liability for its progress because the various disciplines have "different theoretical and methodological orientations and practices which are difficult to reconcile" (ibid.: p.4). A more severe limitation of the research area is the restricted, or even non-existent, empirical basis. Language has emerged from non-language in the distant past and there is no direct trace of the event. Furthermore, the

origin of language is an incident that does not repeat itself and therefore cannot be observed. The remoteness of the event and the lack of direct evidence have generated a multitude of theories and a business of speculating often too vague to allow for specific and detailed empirical predictions, which, according to some (e.g. Cangelosi and Parisi, 2002), are the very essence of science; as it is pointed out, "[...] the study of language origins and evolution remains a somewhat dubious research field" (ibid.: p.4).
It is this problematic aspect of the study of language evolution which computer simulations might help to overcome.

1.2 Computer Simulation of Language Evolution

According to Cangelosi and Parisi (2002), computer simulations are the implementation of a theory in a computer. They are claimed to possess, most prominently, three characteristics:

- The formulation as a computer program entails *explicity, detailedness, consistency*, and *completeness*.
- A theory expressed as a computer program necessarily *generates predictions*, namely, the results derived from the simulation.
- Simulations are not only theories but also *virtual laboratories*, in which researchers can observe phenomena under controlled conditions and test hypotheses and ideas experimentally. Simulations can be repeated *ad libitum*, with varying initial conditions, which is otherwise entirely impossible.

Additionally, computer simulations encourage researchers to take a synthetic, bottom-up and constructive approach that permits the study of problems and phenomena that are analytically intractable, such as those of complex and non-linear systems - of which language and evolution are typical exemplars. It is by virtue of these characteristics that computer simulations by-pass some of the shortcomings exhibited by traditional theorising about language evolution, making them a valuable tool for verifying and assisting

abstract reasoning[1].

What kind of studies have been undertaken with the help of computer simulations? In the discipline's dawning hours, the emergence of shared speech systems (lexicons) has mostly been explored (e.g. Hutchins and Hazlehurst, 1995; Cangelosi, 1999), while in more recent research the auto-organization of various aspects of syntax (e.g. compositionality, recursiveness) has been under investigation (e.g. Kirby and Hurford, 2002; Christiansen et al., 2002; Kirby, 2001), aiming, among other things, at (in-)validating certain tenets held by the post-Chomskyan tradition of syntactical research. Other authors have concentrated on the interaction between linguistic abilities and other behavioral and cognitive skills, such as the the interdependence between the origins of language and motor skills, or the interdependence of language and thought (e.g. Steels, 2002). Among the simulation 'environments' employed were such diverse data structures as artificial neural networks, genetic algorithms, the use of symbolic formal (context-free) grammars, and even game theoretic approaches.

1.3 The model at issue here

The model employed in this thesis to simulate the evolution of language is inspired by the work done in Livingstone (2002) and Livingstone and Fyfe (1999). They have used artificial neural networks and a particular simple learning rule to simulate the emergence of shared lexicons among locally distributed language users - in other words, they have studied the emergence of linguistic dialects. Their goal was to show that the emphasis placed upon social factors as initiators of language diversity in more recent linguistic thought (e.g. Labov, 1972) was disproportionate. They wanted to show that variation

[1]Of course, computer simulations also have limitations, such as a requirement for computational tractability, which necessitates (possibly unduly) simplifications and a difficulty of external validation of the results. One of the profoundest shortcomings is, in my view, that the results of any simulation can only have *negative* character; if anything cannot be derived from the assumptions represented in the computer program then it can be concluded that these assumptions were insufficient *in realitas* to produce the desired result. If, on the other hand, a certain linguistic aspect, say recursional syntax, emerges during the course of the simulation, then the generalization value of the simulation result will depend upon the correctness of the assumptions. It cannot be 'proven' that human evolutionary history has actually taken the path suggested by the simulation.

could result from geography alone. In doing so, they neglected a thorough analysis of the potentialities and impacts of the components of their work, and most of the quantifications made were based upon 'impressions' rather than upon numerical data generated by their results. The current work will add on here; a major part is devoted to the analysis of the various parameters available in the model of Livingstone (2002), and Livingstone and Fyfe (1999) and the role these play in the emergence of language and/or diversity. Aspects of language change will naturally be of concern in a study of language variation. However, the current work tries to go one step further; in addition to 'analysis', 'synthesis' shall be conducted; the second half will be devoted to the emergence of simple (regular) grammars and the emergence of simple morphology, within the current framework. It will be tested whether syntax can already be obtained with very general prerequisites, renouncing on, for instance, complex innate biases. Finally, because the author of this work rejects the view that *either* geography *or* sociology must be held exclusively liable for dialects to come into existence, sociolinguistic factors of language diversity will be incorporated into the model's assumptions. It will be tested if this is possible at all (i.e. whether such factors *can* be simulated within the framework) and if so, whether the results derived by the simulations agree with empirical data and abstract reasoning.

The structure of the work is as follows: Chapter 2 introduces the basic terminology and results of dialectology and language change. Chapter 3 gives an overview of the simulation model employed and the details of implementation. Chapter 4 provides an analysis of the model's parameters and discusses the results observable for the 'basic' implementation. Chapter 5 describes the setup and outcomes of genuinely novel experiments; these include the study of the emergence of simple compositional grammars. Chapter 6 recapitulates the findings obtained and relates these to questions of general linguistic importance.

Chapter 2

Linguistic background

> And the whole earth was of one language, and of one speech. And it came to pass, as they journeyed from the east, that they found a plain in the land of Shinar; and they dwelt there. And they said one to another, Go to, let us make brick, and burn them thoroughly. And they had brick for stone, and slime had they for mortar. And they said, Go to, let us build us a city and a tower, whose top may reach unto heaven; and let us make us a name, lest we be scattered abroad upon the face of the whole earth. And the Lord came down to see the city and the tower, which the children builded. And the Lord said, Behold, the people is one, and they have all one language; and this they begin to do: and now nothing will be restrained from them, which they have imagined to do. Go to, let us go down, and there confound their language, that they may not understand one another's speech (Genesis 11:1-7).

2.1 Dialect and language

Dialectology is the study of dialects, that is, 'subdivisions of a particular language' (Chambers and Trudgill, 1980: p.3). One of its primary theoretical issues is the differentiation of a language and a dialect. How can the two be distinguished? A useful method is to employ the criterion of *mutual intelligibility*, which states that dialects are mutually intelligible and that "a language is a [...] collection of dialects" (ibid.) and that hence, accordingly, if two speakers do not understand each other, they must be speaking different languages; and else, dialects of the same language.

A major problem to this definition arises in the case of (geographical) *dialect continua*, where a 'chain' of lects[1] is spoken throughout an area, characterized by a mutual un-

[1] *Lect*, or *variety*, is used to refer to *any* kind of a language which can be identified in a speech community - whether this be on personal, regional, social, occupational, or other grounds (cf. Crystal,

derstanding of speakers of adjacent regions and a decreasing amount of intelligibility of more distant speakers; people at the two ends of the chain might then find it impossible to understand each other. The challenge involved here is the classification of these lects; where does one language end and where does the next one begin?

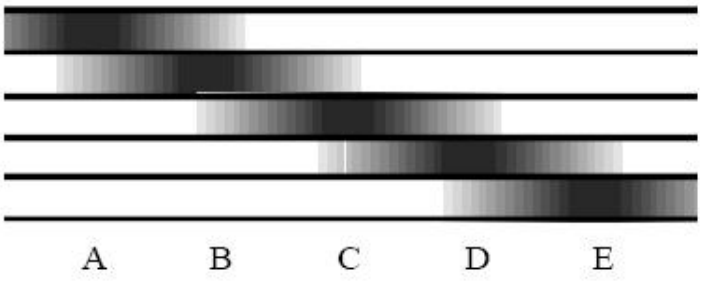

Figure 2.1: A schematic dialect continuum from dialect A to dialect E, showing some degree of mutual intelligibility between adjacent dialects (from Livingstone, 2001).

The phenomenon itself is quite common. For example, an extensive continuum links all the dialects of the languages German, Dutch, and Flemish. Speakers in eastern Switzerland cannot understand speakers in eastern Belgium; but they are linked by a chain of mutually intelligible dialects throughout the Netherlands, Germany, and Austria (cf. Crystal, 1987: p.25). Other chains in Europe include the Scandinavian continuum, linking Norwegian, Swedish, and Danish; the West Romance continuum and the North Slavic continuum (ibid.).

1987: p.25).

2.2 Isoglosses

The term *isogloss* refers to a boundary line separating two regions disagreeing on some linguistic feature. An example is the famous isogloss that runs across England, from the Severn to the Wash, distinguishing northern speakers who pronounce a rounded [u] in words like *cut* from southern speakers who keep the vowel open and unrounded.

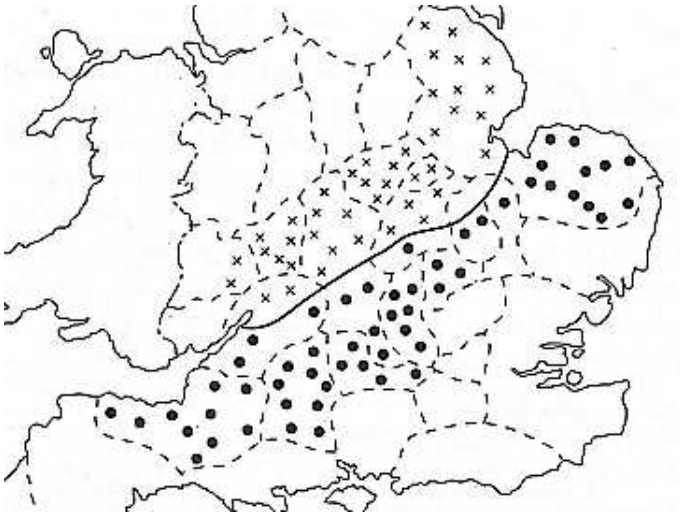

Figure 2.2: The line on the map of southern England separates the area in which the vowel in a word such as *cut* is [a] (black circles) from the area in which the vowel is [u] (crosses) (from http://www.phon.ucl.ac.uk/home/johnm/sid/isogloss.htm).

Term	Separates	Examples
isolex	lexical items	*munch* vs. *muncheon*
isomorph	morphological features	*dived* vs. *dove*
isophone	phonological features	*put* /put/ vs. /pat/
isoseme	semantic features	*dinner* (mid-day meal) vs. (evening meal)

Table 2.1: Structural categories of isoglosses (after Crystal, 1987: p.28).

When the concept of isoglosses was first introduced in 1892 by the Latvian dialectologist Bielenstein, it was supposed that they would provide a clear method for identifying dialect areas. It was assumed that the isoglosses for many linguistic features would coincide, demarcating one dialect from another (ibid.: p. 28). However, the only prominent

and recurrent *pattern* found seemed to be the absence of pattern, an almost chaotic criss-crossing of isoglosses. Only when viewed from a more distant scale, distinct dialect areas could be determined, constituting the fact that, while isoglosses rarely coincided, they did often run in the same general direction. Some areas, called *focal areas*, were seen to be relatively homogeneous, containing few isoglosses. Where they merged, the number of isoglosses and hence linguistic variation increased; these regions became known as *transition areas*.

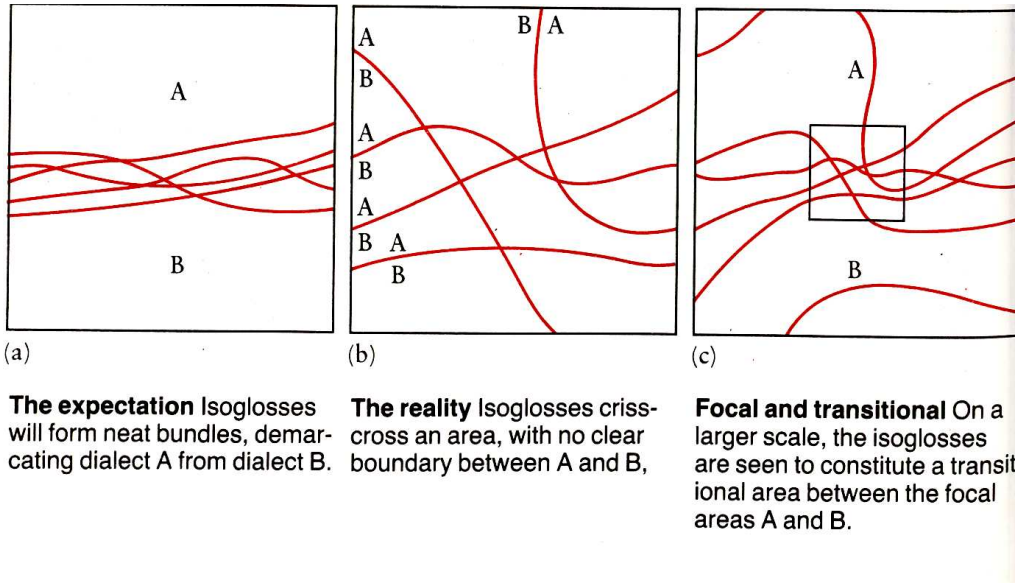

Figure 2.3: Isoglosses: focal and transitional areas (from Crystal, 1987: p.28)

As Chambers and Trudgill (1980) point out, the pattern of criss-crossing (on the small scale) isoglosses apparently describing a chaotic variety of dialect feature combinations is today recognized as a typical pattern for any region that has a long settlement history. Another pattern distinguished by dialectologists is a particular isogloss's delimiting of areas in more than one part, without continuity. This means that a linguistic feature exists in two (or more) parts of the region, while those parts are separated from one another by an area in which a different, opposing, feature occurs. Such a pattern is assumed to reflect a late stage in the displacement of a formerly widespread linguistic feature by an innovation - in earlier times, the feature now found in the isolated areas was also present in the in-between areas. Its status is now that of a *relic feature*.

Figure 2.4: The discontinuous distribution of the rhotic areas (as of 1950) on this map of England indicates that 'rhotic' is a relic feature in England (from http://en.wikipedia.org/wiki/Rhotic_and_non-rhotic_accents)

2.3 Sociolinguistic approaches to dialectology

Unlike traditional dialectology, which focussed solely on the relationship between language and geography and on the spatial differentiation of language, modern dialectology has also explored the interplay of language and social features; "modern [...] dialectologists take account of socioeconomic status, using such indicators as occupation, income, or education, alongside age and sex" (ibid.: p.32).

A good example of the effect of social class on language is the dropping of the *h* in English. In British English, the most prestiguous accent[2] pronounces /h/ at the beginnings of words such as *head*, whereas it is common in most other dialects of England and Wales to omit, or drop, the phoneme in initial position. The results of two studies of this variable[3] carried out in Norwich and Bradford are shown below.

As can be seen, /h/-dropping correlates with social class - increasing as one moves down the scale. As one can also see, the proportion is always greater in Bradford, suggesting that the phenomenon has been longer established in that area. Finally, the example

[2] *Received Pronunciation.*

[3] A *linguistic variable* is a unit with at least two variant forms, the choice of which depends on other factors such as sex, age, social status, and situation. For example, in New York, (r) is a variable, with the two forms, /r/ and zero (as when speakers sometimes pronounce it in words like *car* and sometimes do not) (cf. Crystal, 1987: p.32).

Class	Bradford (in %)	Norwich (in %)
Middle middle	12	6
Lower middle	28	14
Upper working	67	40
Middle working	89	60
Lower working	93	60

Table 2.2: *h* dropping in two British cities (after Chambers and Trudgill, 1980: p.69)

illustrates that the relationship between dialect and social class is a matter of degree, tendencies, and probabilities; it is not the case that some groups use one variant and others the other; rather, all groups use both variants, but in different proportions.
Other aspects examined by modern dialectologists include the relationship between sex and language, the relationship between ethnicity and language (e.g. the English of black, as opposed to white, speakers), and the relationship between situation and language (formal style vs. casual style).

Why do social factors influence language behavior? First, people are naturally more affected, and in particular linguistically, by the members of the social network to which they belong - if it be only because they spend more time, and hence (linguistic) interaction, with them. Second, a distinct language variant provides the basis for a common identity[4]. Finally, a separate lect might be prescribed by social status and role, as in the case of ceremonial language (Roman Catholic liturgy used to be held in Latin) or in the Hindu caste system, where certain variants of language are exclusively reserved for Brahmin (highest social class) speakers (cf. Crystal, 1987, p: 41).

2.4 Language change

Another important aspect of language variation is language change. Why do languages change? In the literature, several arguments have been put forward, such as (ibid.: p.333),

[4]As in the case of language change observed on Martha's Vineyard where, according to Labov, local fisherman subconsciously adapted their pronunciation of the vowels [ai] and [au] in order to distinguish themselves from the massive influx of tourists (cf. Crystal, 1987: p.332).

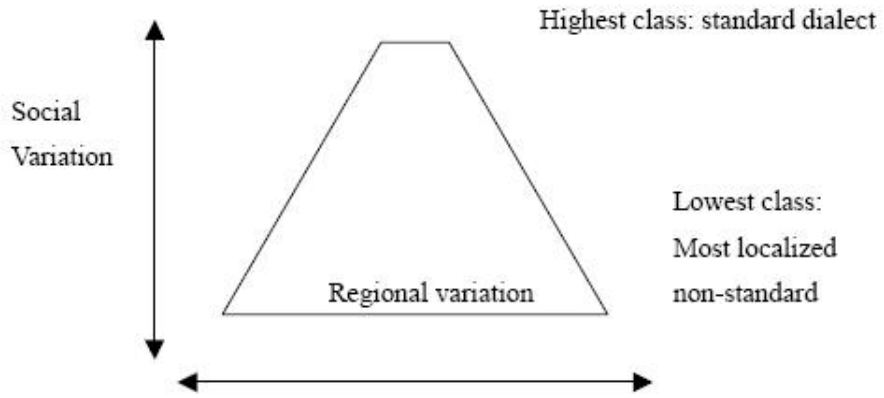

Figure 2.5: The relationship of geographical and social factors in determining dialect. Empirical studies in countries such as Great Britain have shown that there is an inverse correlation between the effects of social and regional 'location' upon language: the higher a person is on the social scale, the less her dialect is conditioned by the region in which she lives. Among the members of the lowest social class the widest range of local accents and dialects is encountered (from http://www.cels.bham.ac.uk/resources/essays/macedo6.pdf).

- **geography**. Languages change because language users move away from one another or because they come into contact with one another. This latter fact also illustrates that languages might change (merely) as a result of the existence of dialects, when speakers can influence each other.

- **physical world**. Things in the cultural/physical world may change (such as the invention of new objects or the abandoning of old customs), requiring an 'updating' of lexicons.

- **imperfect learning**. According to this view, children are the initiators of language change. They might learn the adult forms imperfectly, and a new standard could gradually emerge.

- **social prestige**. People might adapt their speech to the speech of those they 'admire'. This process may be conscious or subconscious. There might be a movement towards a favored accent or dialect.

Other possible factors of language change concern the nature of language itself rather than extra-linguistic causations. Among these, *ease of articulation* - for example, a principle of least articulatory effort might preclude certain sound combinations to occur so that sound systems should become simpler over time - and *analogy* - the trend towards a regularisation of irregular grammatical patterns - have been mentioned.
On the other hand, such 'hermeneutic' (or functional) reasoning about the origin of language change is rejected by some researchers, such as Lass (1997). They claim that such explanations are misleading because they are uninformative. As an example, the analogical levelling of the *i*-umlaut in consonant noun-stems in historical English is given. In Old English, some nouns such as *fōt* 'foot' had, for various historical reasons, two stem-allomorphs; nominative, accusative, genetive singular, and genetive and dative plural had a non-umlauted vowel while the rest of system had an umlaut, so that forms like *fōt* (nom. sg.) and *fēt* (nom. pl.) were observed. Hence, umlaut in these paradigms showed 'purposeless variety', explicable only as historical residue; synchronically, it correlated with neither number nor case, but with seemingly arbitrary combinations (cf. Lass, 1997: p.342). According to Lass, a hermeneutic approach to language variation should predict that such a functionless pattern of language use be remedied. This was indeed the case for English where, by the Middle English period, the umlaut variation had become a marker of number (i.e. singular nouns had the non-umlauted vowel, plural nouns the umlauted), and hence lost its 'purposelessness'. Still, Lass claims that the interpretation is at issue. In a related language like Icelandic, for example, such a move towards a 'desirable' state has not occurred[5]. Lass summarizes the problem when invoking such a hermeneutic principle (which he calls, in the example, the MSPV principle) for the explanation of language change and concludes,

[5]Instead, forms such as *fœt-i* (dat sg) co-occur with *fót-a* (gen pl).

> The problem should now be clear. If we invoke MSPV as an explanation only for good outcomes, and allow bad ones to be not counterexamples but simply non-instantiations of something 'tendential' in the first place [...], the MSPV explanation is invincible, and therefore uninformative. One can't really say that 'the mind shuns purposeless variety' and use that as an explanation for all cases where it apparently does, because so often it seems not to shun it at all (ibid.: p.344).

He goes on to say that the (*a priori*) probability of change, with respect to a certain feature, is exactly one half (either the feature will change or it will not), and that thus, "One could predict results of this kind even if there were no motivation at all for change, but only change itself as an axiom [...] Change could in principle, given the data we have, be random rather than motivated, and the results would probably look pretty much the same" (ibid.: p.346).

This statement also points to the direction of Lass's explanation for language change. In his view, it is enabled by the presence of (linguistic) redundancy[6] and *randomly* selects its course, "Semiotic systems are full of noise and redundancy and junk, and this non-functional, non-semiotic debris is crucial for change, and may indeed be the main thing that makes it possible" (ibid.: p.310). Redundancy can act as a 'placeholder' wherein change can emerge, and junk can be recycled. Again, he warns not to prematurely attach 'meaning' to the randomness exhibited by such variation, "Since outside of these narrow boundaries all language states are equifunctional, change cannot 'improve' a language state, or 'meet needs' that are not already meet" (ibid.: p.366).

Of course, it cannot be hoped to settle the question of whether language change is *exclusively* functional or random or to "prove" that some factor *alone* is responsible for its presence; but, what can be done at least is to test the plausibility of some of these alleged originators of language variation and to analyze their respective importance, for example by switching on some of the parameters mentioned while disabling others. A discussion of these issues will ensue in chapters 4 and 5.

[6] One example of such redundancy being the (useless) vowel alternation discussed above.

Chapter 3

Implementational Details

> In its most general form, a neural network is a machine that is designed to model the way in which the brain performs a particular task or function of interest; [...] (Haykin, 1994: p.2)

Disregarding the success of models relying on symbolic approaches[1] to the simulation of language evolution, the simulation model under analysis here employs artificial neural networks (ANNs) to obtain insight into the subject matter. Although symbolic approaches are often more accessible for the human interpreter, and in particular allow the modern linguist to translate her conceptualizations more comfortably, ANNs are the method of choice in this simulation for three reasons. Firstly, they represent - as suggested by the quotation at the beginning of this chapter - a data structure devised in close analogy to the workings of the human brain, and can thus be regarded as more 'natural', for example, than the type of generative grammars employed by the post-Chomskyan generation of syntacticians. Secondly, the simulation model made use of in this work is characterized by its simplicity, thus allowing the researcher to concentrate more deeply on the actual topic of interest - the complexity is to be found in the evolution of language, not in the data structures. Thirdly, using ANNs for the simulation of language evolution enables the employment of thoroughly-studied methods and algorithms, as a result of almost forty years of research. No *ad hoc* solutions have to be tailored.

[1] Most of these approaches are based on the rule-based evolution of context-free grammars by a set of artificial agents (for example Kirby and Hurford, 2002; Christansen et al., 2002).

3.1 A simple model for the evolution of language

The research project presented in this paper is based on the simulation of the linguistic behavior of a *population* of individuals, referred to as *agents*. These agents are capable of acquiring and speaking a very simplified language in response to arbitrary, a priorily defined *meanings*[2] that are common to all agents. A speaker agent produces language tokens (a *signal*) to denote the current meaning and a listener agent attempts to interpret what the original meaning was from the received signal. Each agent participates in many interactions, sometimes as signal producer, sometimes as listener. The goal is the emergence of (*locally* or globally) *shared* lexicons or languages.

As indicated above, an agent is implemented by a fully connected ANN consisting of two layers of nodes, an *input layer* to hold the 'meanings' which the function realized by the connecting weights maps to a signal on the *output layer*. Each unit ('neuron') in each layer can take on two values, implemented as $(+1)$ and (-1) respectively, henceforth sometimes referred to as $(+1)$ and (0) for the sake of readability; sometimes, the two values may also be interpreted entirely differently in the course of this work, for example as grammatical characters of the alphabet $\Sigma = \{a, b\}$.
In the model's default operational mode (as adapted from Livingstone, 2001), an agent's neural network comprises an input layer consisting of three neurons plus a bias node[3], and three output neurons. Output units may take on each of the two values so that agents are capable of uttering $2^3 = 8$ distinct signals. One of the great advantages of this signal representation is the ease with which signals can be represented as color pixels, allowing the investigator to *visualize* an agent's language, the benefits of which, the author hopes, will become clear in the course of this work. In order to provide redundancy in the representation of meaning[4], the meaning set is limited to the three meanings $\{100, 010, 001\}$; in other words, agents try to discriminate between 3 different

[2] In Saussurean terms, *signifiés*.

[3] The value of the bias node is consistently set to $(+1)$. On the role of the bias neuron, see Haykin (1994), and Chapter 4.

[4] The role of which will be investigated in Chapter 4. See also the discussion on language change in the previous chapter.

meanings with the help of 8 different signals.

When a communication between two agents is scheduled to take place, one agent takes turn as a speaker (or, *teacher*), the other as a listener (or, *learner*). One of the meanings common to all agents is randomly chosen, presented to the teacher's input neurons, and fed forward through the ANN to generate the teacher's signal for that meaning. The *signum* function thresholds each output unit to a binary value. The learner's ANN is then 'reversed' by presenting the speaker's signal to the listener's signal (i.e. output) layer and feeding this activation back to the meaning (i.e. input) layer, where either a winner-take-all competition determines the meaning understood by the learner (in default mode), or, again, a thresholding function activates the input neurons[5]. In order to provide for adaptation and the emergence of shared language, the learner updates the weights of his net, using the updating rule

$$\Delta w_{ij} = \eta \cdot (x_i - x'_i) \cdot y_j,$$

where $\eta \in (0, 1)$ is the learning rate parameter, specifying the speed (or willingness) of adaptation to the teacher, $x = (x_1, x_2, x_3)$ denotes the original meaning the teacher was asked to verbalize and $x' = (x'_1, x'_2, x'_3)$ the meaning understood by the learner (the *generated meaning*); y_j is the value of signal neuron j. As shown, learning occurs only when the learner misclassifies the signal, for otherwise $x_i - x'_i = 0$. One important implication of this rule is that language bargaining is seen as a matter of the listener *only*. The weights of the speaker's net are retained unmodified, irrespective of the listener's success of understanding the teacher.

[5]In the default mode, the meaning set is given by $\{100, 010, 001\}$ (see above), and thus, a winner-take-all competition over the meaning neurons has to be conducted in order for the listening agent to determine the original meaning; put differently, the listener identifies the meaning neuron that has greatest activation and sets this neuron to (+1), and the rest to (0).

3.2 What does the updating rule do?

Suppose that the original meaning $m \in \{+1, -1\}^k$ has a value of $(+1)$ at position i, where the meaning generated (by the learner) for the uttered signal s has a (-1) at i. Then $\Delta w_{ij} > 0 \iff y_j > 0$, which means that the connection between input neuron i and output neuron j is increased if $y_j = 1$, and decreased if $y_j = -1$. Suppose $y_j = (+1)$. Then after the net update, if the learner is again prompted to generate a meaning for a signal having a $(+1)$ at position j, this stronger connection between input neuron i and output neuron j will make it more probable that the generated meaning will be positive at position i - as it was the case in the original meaning. If, on the other hand, $y_j = (-1)$, the connection will be weakened, $\Delta w_{ij} < 0$, thus contributing little (of the negative value y_j) to the induced field[6] of input neuron i, again increasing the probability that meaning neuron i will be positive. In the case where $x_i = -1$ and $x'_i = +1$, the argument is analogous. To conclude, after the net update the learner will, in particular, be more inclined to interpret the signal s as the original meaning m[7].

As an example, suppose a learner has two input and one output neuron, with weight values $w_{11} = 0.4$ and $w_{21} = 0.6$. Suppose, the teacher is asked to verbalize the meaning $(1, -1)$ and does so by uttering 1. Then the meaning generated by the learner (in response to the signal 1) will be $(-1, 1)$ (in default mode; the second input neuron is the 'winner'), so that the net update for the learner will be $\Delta w_{11} = \eta \cdot (1 - (-1)) \cdot 1 = 2\eta$ and $\Delta w_{21} = \eta \cdot (-1 - 1) \cdot 1 = -2\eta$. If, say, $\eta = 0.1$, then the revised weight values will be

$$w'_{11} = w_{11} + \Delta w_{11} = 0.6, \quad w'_{21} = w_{21} + \Delta w_{21} = 0.4,$$

[6] See Haykin (1994): p.11.

[7] Although this makes it reasonable why a particular learner should adapt his signal-meaning associations to that of a particular teacher, it is not clear at all (*a priori*) that this simple updating rule should effect the emergence of shared linguistic behavior - if indeed it does - over a whole population of agents and many meanings to be interpreted. It might happen that random and disorderly meaning-signal mappings arise among agents where one lexical entry learned from one agent is immediately annihilated by what is learned from another. If indeed the model enables the acquisition of shared language, this should be called *a miracle of language acquisition.*

so that if the learner is again summoned to decode the signal 1, he will give it the same interpretation as the teacher.

3.3 Spatial organization of agents

Since a major part of this work is dedicated to the analysis of the model suggested by Livingstone (2001), and Livingstone and Fyfe (1999), the study of the emergence of dialects will be of primary concern. For this sake, agents are situated in a one dimensional world - they are placed in a row with unconnected ends, so that each ('inner') agent has two immediate neighbors. Communication between agents is limited by pre-defined neighborhoods based on distance; the chance that two agents will talk to one another is determined by a normally distributed curve centered on the listener. This creates a population structure where communication is localized but which does not have explicit group boundaries. However, exceptions from this default operational setting will be investigated, too, as when communication between agents is geographically unbounded or when explicit group boundaries are introduced in the aftermath of the examination of sociolinguistic factors of language variation.

Agents additionally pass through a life cycle consisting of two stages, child and adult. Children and adults populate separate rows of identical size. The adults' only task is the transmission of the language they have acquired in their youth (i.e. adults function only as teachers); child agents may both act as learners and teachers (for other children). Teachers are selected from their respective rows according to the current neighborhood size, relative to the learner. After training, the existing adults are removed, the current child population is aged and a new row of children populated. The set of agents jointly existing in one of the two life stages is referred to as *generation*; the sequence of generations chronicled during one evolutionary process as *civilization*.

To summarize, the basic algorithm operated in the default mode looks like follows (cf. Livingstone and Fyfe, 1999). The first generation of agents is initialized with random weight values and then provide training for one another for t training rounds (or,

communicative episodes),

```
For each agent (in random order):
    Pick another agent to be the teacher according to neighborhood size
    Randomly pick a meaning, generate signal from teacher for meaning
    Train pupil on meaning and signal.
```

Each successive generation is initialized with either uniform or random weights. These new agents may then learn from their parents and/or fellow child learners.

Chapter 4

Analysis

4.1 Introduction

> [...] the appearance of purpose is an illusion: in the historical perspective, there's nobody to have any purposes [...] and hence neither purpose nor purposelessness, but simply arbitrary and random variation (Lass, 1997: p.351.)

The computational model of language evolution presented in the last chapter depends on various parameters. The overview over these below gives an outline of the types of questions that will be addressed in the remainder of this chapter.

- **Neighborhood size**. By changing the standard deviation of the normal curve used to determine the likelihood of a neighbor's selection, the effective neighborhood size can be varied from one that only includes the immediate neighbors to one that includes the entire population. With the standard deviation large enough, a scenario can be created analogous to a selecting of neighbors from a uniform distribution.
  In general, if data is normally distributed, about 68% of the values are within 1 standard deviation away from the mean, about 95% of the values are within two standard deviations and about 99.7% lie within three standard deviations. For the model under investigation this means that if a standard deviation of, say, 1 is chosen, the probability that the left or right neighbor (or the agent itself, if this is allowed) of a given learner is selected as a teacher in a communicative act is more than 2/3.

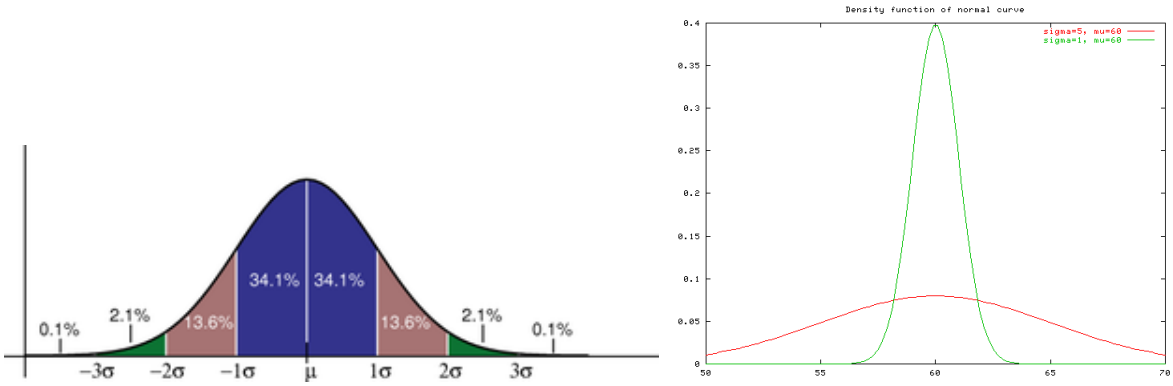

Figure 4.1: Left: Probability mass as a function of standard deviations (from: http://en.wikipedia.org/wiki/Normal_distribution), right: Normal density function with different standard deviations ($\sigma = 1$ (green steep curve) resp. $\sigma = 5$ (red flat curve))

Additionally, a gradual increase (or decrease) of the neighborhood size could be employed to simulate a varying degree of mobility amongst agents; for example, an increasing neighborhood size over time could be interpreted as being a result of technological progress, and thus the effect of technology upon language could be investigated.

- **Initialization**. The first generation of agents has no previous generation to learn from, and so must be initialized somehow; the weights of the net representing a particular agent must be set. Likewise, each child generation needs to be initialized. It is possible to initialize the population in such a way that all agents use the same signaling scheme, e.g. by setting the weights of the population uniformly; alternatively, the weights may be set randomly[1].

- **Learning rate**. The learning rate parameter $\eta \in (0, 1)$ may be varied in order to study its effect on the number of dialects. A small value of η would correlate with conservative, "hesitant" learning, while a larger value might be interpreted as 'quicker' comprehension, higher intelligence, etc. but also as additional factor of noise (particularly for $\eta \approxeq 1$). Since the learning rate could also be taken as the

[1]While the first variant could be regarded as an analogy of either a *tabula rasa* interpretation of human development, or of uniformly biased brains, the second would account for (an assumed) individuality of the infant human.

hearer's "confidence in the speaker" or as the speaker's status (as attributed by the hearer), η can be used additionally to implement a social function.

- **Number of signals and meanings**. By default, agents are 'equipped' with three different meanings (the agents' cognitive world) to which they can assign 8 different symbols. Besides the testing of the use of just 'more' possible meanings and/or signals, various relations between the two factors may be explored, particularly, in order to test the hypothesis that linguistic change is explained in part by redundancy.

- **Topology.** Agents are by default arranged in a one dimensional world with unconnected ends. What happens if they are situated in higher dimensions? Will the results be similar or will qualitatively different outcomes be observed?

- **Training**. Each generation is 'trained' by its parent-generation (except, of course, for the first). Additionally, 'inter-child' training may (or may not) occur; in other words, after having been trained upon their parents, agents may (or may not) talk to each other. It might also be possible to devise more realistic linguistic scenarios where child agents might alternatingly speak with adults and fellows.

- **Number of communicative episodes per generation**. Each agent in each child generation participates in linguistic interaction in t communicative episodes (see previous chapter); t determines the amount of training (or exposure to language) that each agent receives. What is an optimal value of t? What is the minimally required value of t?

- **Non-uniform distribution of meanings**. In the default setting, all meanings are drawn with equal probability to be the subject of a given conversation. What happens when topics are chosen on the basis of non-uniform distributions? Will the lexical entries for higher frequency meanings have a different form (in whatever respect) than the entries for low frequency meanings?

- **Noise**. By default, the communication between agents is noise free and perfect

replication of signaling schemes across generations is possible - something that is not possible in human language learning because, among others, the linguistic evidence presented to children is insufficient to allow perfect replication (cf. Livingstone, 2002). One of the issues that might be addressed is whether (random) errors in language use or comprehension can help drive language change.

4.2 An introductory example

The following example serves as an introduction to the experiments that will be conducted in the succeeding sections. Here, and henceforward, the population consists of 120 agents. Each agent in a generation was trained for $t = 80$ rounds by the agents of the parent generation; the first generation was randomly initialized and each agent there likewise performed in 80 communicative acts as a listener; an agent could not be its own teacher. There was no inter-child training, that is, after the first generation, agents would only learn from their parent generation. All individuals of one generation later than the first were uniformly initialized, in that the weights between their input and output neurons were set to zero; hence, they all used the same mapping between meanings and forms (all meanings were mapped to the signal 111). The learning rate parameter η was set to a moderate 0.1 and the standard deviation of the normal distribution used to determine the neighborhood size to a (relatively large) value of 2.5. 1000 generations were chronicled. Agents had to negotiate a signaling scheme for the three meanings 100, 010, and 001. A signal consisted of three bits, and thus, the agents had $2^3 = 8$ possible signals at their disposal. Using each of the three bits to set one of the colors red, green, and blue, the maps shown in Figure 4.2 were obtained by prompting each agent in each generation for a given meaning, after training had been completed. The agent would respond with its signal, which could then be plotted on the map. Horizontally, the pixels (signals) of the agents in the same generation are displayed. The vertical axis is the concatenation of the 1000 generations. The advantage of the representation of the signals as pixels is not only the visualization, which allows the comparison of dialects across many hundreds

of generations; the visualization itself has the advantage that similarity of color stands for similarity of signal. A dark blue color next to a light blue may indicate mutually intelligible dialects.

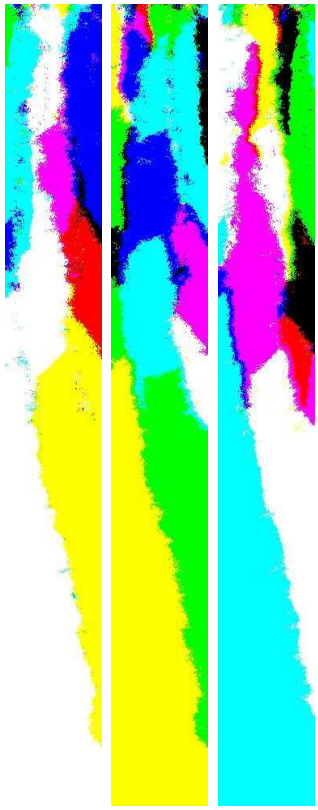

Figure 4.2: $\eta = 0.1$, $\sigma = 2.5$. Signals for the meanings 100, 010, and 001 (left to right). Horizontally, the signals of the 120 agents in each generation - represented as color pixels - are displayed (for each meaning). The vertical axis shows the different generations, 1000 altogether.

What can be deduced from these maps? It can be seen that for each meaning connected dialect regions have emerged in each generation, and that these dialects evolve over time, gaining and losing speakers; sometimes a dialect extinguishes, and another one replaces it. Likewise, the maps clearly illustrate that the same speakers use different signals to convey each meaning. For example, considering only the last few hundred generations, the speakers on the left part of the map associate a signal represented by the color white with the first meaning, a 'yellow' signal with the second, and a light blue with the third. Next, within the populations dialect continua form whereby chains of -

one might conjecture - mutually intelligible signal schemes span across the population, linking the various dialects. Between distinct dialects, the point where signal use changes tends to be different for each meaning (as can be recognized primarily in the first few hundred generations), which illustrates that indistinct boundaries between the dialects have formed; put differently, isoglosses display the pattern of criss-crossing similar to the scheme observable in human dialects (see Chapter 2.2).
What else can be seen? For each meaning, there are relatively few dialects spoken among the agents: 2 to 4 in the first hundred generations, then two dialects, and finally a single global lect has emerged. It may be assumed that this development of increasing uniformity (as emphasized by the decrease of entropy; see Figure 4.3) is mainly due to the large value of the neighborhood size, a question investigated in the following sections. Additionally, individual dialects seem to be very homogeneous, meaning that among the speakers of a particular dialect few idiolect variants may be discerned. A tentative conjecture would attribute this fact to the absence of noise in the communication acts and to the uniform signaling system that all members of the child generation are initialized with.

Next, a short analysis of the agents' communicative behavior and their communicative successes as a function of generation age[2] will be given. It was mentioned earlier that each agent in each generation is exposed to t (= 80, in this example) communicative acts as a learner; a neighboring (parent) teacher is selected, and the learner adapts its weights in order to conform with the teacher's signaling scheme. Given that the number of agents in the simulation is 120, in each of the t communicative episodes a score of maximally 120 successes (i.e. speaker and listener map a randomly chosen meaning to the same signal) can be achieved. Considering that the agents adapt to each other over time, one might presume that the number of successes will increase with communicative episodes.
Figure 4.4, where the communicative successes of generation 1 (the initial generation) and generation 400 are compared, suggests that this is actually the case. With no structured communication scheme existing for each generation of agents prior to training, the number

[2]By 'age', I mean the distinction between 'early' and 'late' generations. In this sense, generation, say, 17 is older than generation 11, 3, or 1.

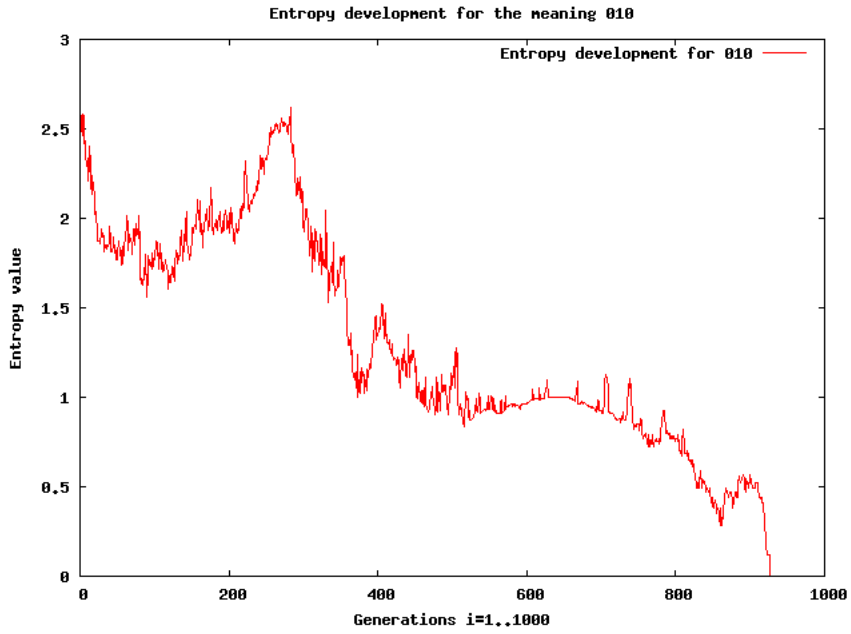

Figure 4.3: Decline of entropy for one of the three meanings (for the others, a similar development can be observed). The decline illustrates the increasing uniformity of the form-meaning mappings among the agents.

of successes amounts to guessing at this point (a value of $\approxeq 0.3$ in the sample; note that there are three possible meanings). It can be seen that the first generation gradually - but not monotonically - evolves a locally shared communication system, and finally, within their local boundaries, agents are capable of mutual comprehension in approximately 60 to 70% of the time - a doubling of performance after 80 episodes of communication. Generation 400 is not only much more successful (with a final score of around 95%) but also learns much more rapidly. After 5 episodes, the agents' performance in mutual understanding has more than doubled, and after less than 10 episodes, it is already well beyond 90%, a level it will not fall below any more.

Why is this? The agents of the first generation initially - prior to training - manifest chaotic linguistic behavior[3]; forms and meanings are randomly correlated. What an agent then learns from its right neighbor may be undone by its left neighbor, and so successful communications are rather accidental. Only gradually do the members of this

[3]Note that these agents are randomly initialized.

population agree (again, not globally, but locally) on shared lexicons. Agents of later generations are in a much more convenient situation. Initially, they are likewise unable to understand their parents. But for each agent, the parents in its neighborhood have already converged to shared behavior. What an agent then learns from its right neighbor may only be amplified by its left neighbor.

As illustrated in Figure 4.5, the overall degree and speed of communicative success is still increasing with generation age, yet the discrepancy between late generations is clearly not so striking as the one between the initial generation and posterior ones (and it may be suggested that after a certain threshold - with respect to generation age - has been exceeded, all differences in speed and quality will be equated).

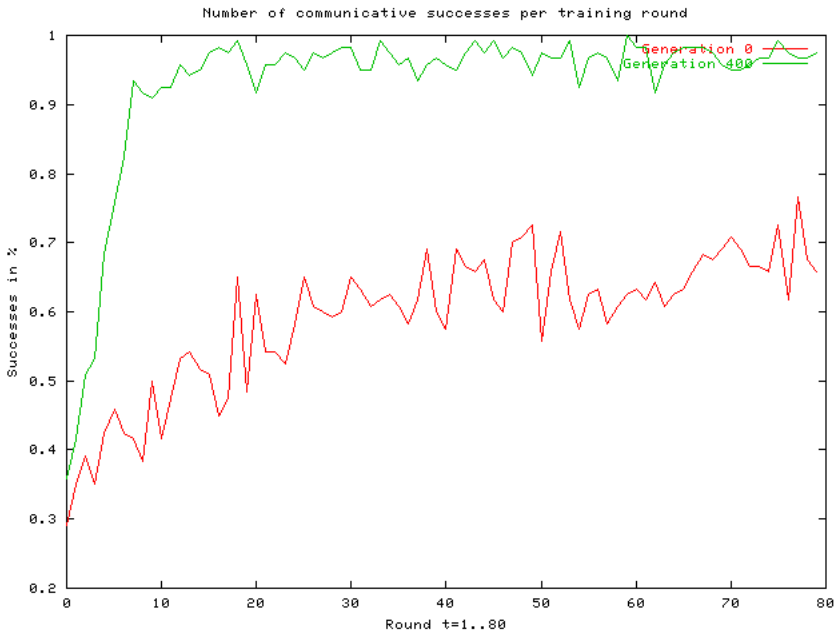

Figure 4.4: Comparing the number of communicative successes (in %) between neighboring agents as a function of communicative episodes. Displayed are the first generation and generation 400. In both cases, mutual intelligibility is increasing with communicative rounds, yet generation 400 is both quicker and more successful.

To conclude this section, a word must be said on the agents' development of distinctive signals for each meaning. Because in the examples discussed so far there are always more (possible) signals than there are (actual) meanings, the mappings from signals to

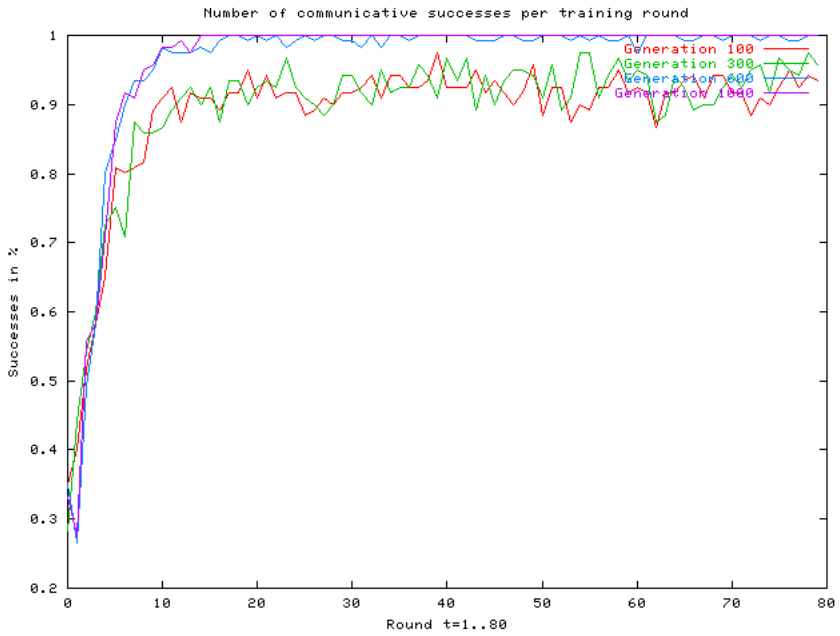

Figure 4.5: Comparing generations 100, 300, 600, and 1000. The difference between the generations is clearly decreasing. Still, generations 600 and 1000 are at a discernible 'advantage' over the preceding two generations.

meanings cannot be one-to-one[4]. The functions (represented by an agent's neural network) mapping from meanings to signals, on the other hand, allow to be examined for this property. What then is expected to be found (and the graphical evidence examined so far does support this belief) is that distinctiveness in signal use increases with communicative episodes[5]. An explanation goes like follows.

When a new generation later than first is 'born', individuals are initialized with all weights set to zero, which means that they will assign to each input (meaning) the same output (signal), namely 111[6]. Therefore, an initial degree of (signal) distinctiveness, defined as the number of signals per meaning, of $1/n$ is observed, where n is the number of meanings supplied to the agents. When many generations have already passed and (parent) agents have already learned to distinguish meanings via signals, children will now do likewise and distinct signals for each meaning will soon establish among the children, merely as a

[4]That is, injective.
[5]And hence, the mappings become 'more' injective.
[6]Because $\operatorname{sgn}(0) = 1$.

result of children imitating their parents. However, why would the very first generation learn to differentiate?

As the maps considered so far have illustrated, diversity is unable to return to the system of dialects once agents have agreed on a global dialect - a globally shared communication network - in the absence of noise. Therefore agents would not *learn* to use different signals for different meanings, if the first generation were uniformly initialized (unless this uniform initialization would entail distinctiveness, in which case, again, it were not *learned*). The first aspect to note is therefore that this generation's random initialization is a necessary prerequisite to signal distinctiveness. Imagine now that there are only two agents, two meanings (say, $(1, -1)$ and $(-1, 1)$) and two possible signals. Assume that agent A maps both meanings on the same signal and agent B injectively assigns meanings to signals. When agent A, acting as a teacher, is asked to present a signal for the first meaning, agent B, as a learner, will adapt its weights in order to correct its verbal behavior; i.e. the weight vector of agent B will be moved in a certain direction. If, at some later point, agent A is acting as a speaker again, this time being asked to verbalize meaning 2, i.e. $(-1, 1)$, agent B may have to correct its weight vector again - but to move it in the *opposite* direction, thus undoing what has previously been learnt. When agent B is acting as a teacher, no such contradictory impetus is exerted. Agent A is thus less successful in conveying its language code than agent B. There is no central control, no optimization strategy for finding an adequate communication system, only random processes, an 'invisible hand', by which some agents have a slightly more differentiated signaling system at the beginning; and these will over time have a greater chance of being adopted, or learnt, by others[7].

The data supports the conjectured hypotheses (see Figure 4.6). Additionally, as entailed by the argument given above, whereas the first generation performs seemingly random movements in order to 'achieve' a higher degree of signal distinctiveness, genera-

[7]Or, to use another terminology, these meaning-signal mappings will have a greater chance of survival or replication.

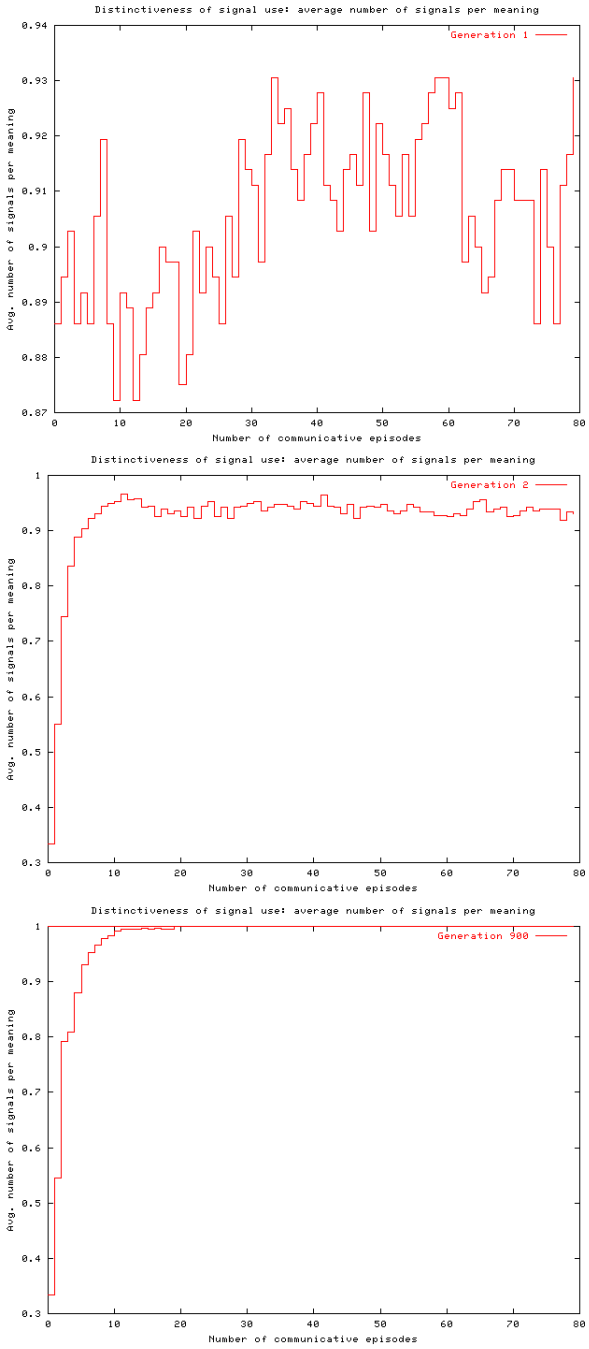

Figure 4.6: Average number of signals per meaning as a function of t after 1,2, and 900 generations (with parameters as in Figure 4.7).

tion 2 is already more successful at this attempt[8] [9], and generation 900 quickly associates

[8]The difference between the first two generations is not so striking as might be suggested by the first impression of the diagrams, however; the first generation has an initial degree of distinctiveness of about

a different signal with each meaning.

4.3 The role of `inter-child training`

As suggested by Livingstone and Fyfe (1999), a child generation may in addition to child-parent conversation engage in 'mutual conversation' where each child agent is acting as a listener for, say, $t/2$ episodes and the speaker, or teacher, is selected among the children, not among the parents. It is stated that this second period of learning should occur *after* children have been 'trained' by their parents.

The maps illustrating this situation (see Figure 4.7) again display (mildly) blurred boundaries between dialects, and a non-identity of boundaries for different meanings, or, equivalently, 'criss-crossing' isoglosses. Approximately 5 to 6 different dialects in the very first generations transform into 3 to 4, and finally collapse into a single global language. This process - which is so unlike the developments known for human languages, where variety is a characterizing feature - now takes place much quicker, however, with total convergence obtained after only about 100 generations. It thus seems that an additional period of inter-child communication helps regularize language performance - a fact that may be attributable primarily to the additional conversational activity.

Indeed, when looking at the maps displaying the dialectal evolution of the language

0.85 and then remains in the interval $[0.85, 0.93]$, whereas the second begins at 0.33, rises rapidly, and then stabilizes at values of about $[0.9, 0.95]$.

[9] Why is the initial degree of signal distinctiveness for the first generation so high ($\cong 0.85$) given that the weights are initialized randomly? Disregarding the bias introduced by the random values assigned to the weights (note that these values are drawn from a uniform distribution from the interval $[-0.5, 0.5]$), it may be assumed that all signals are equally likely to be associated with one of the meanings. Let X_i denote the random variable counting the number of distinct outcomes when throwing a dice with 8 (= number of possible signals) sides 3 (= number of meanings) times. Then $P[X_i = 1] = \frac{1}{8} \cdot \frac{1}{8} = 1/64$, and so $P[X_i > 1] = P[X_i >= 2] = 1 - P[X_i < 2] = 1 - 1/64 = \frac{63}{64}$. Since $P[X_i = 3] = \frac{8 \cdot 7 \cdot 6}{8^3} = \frac{42}{64}$, it follows that $P[X_i = 2] = 63/64 - 42/64 = 21/64$. Therefore, when repeating this triple throwing n times and counting up the respective scores (i.e. $X := \sum_{i=1}^{n} X_i$), one obtains

$$\mathrm{E}[\sum_{i=1}^{n} X_i] = n \times \mathrm{E}[X_i] = n \times (\frac{1}{64} + 2 \cdot \frac{21}{64} + 3 \cdot \frac{21}{32}),$$

and hence $\mathrm{E}[X] = 316.875$, yielding a score of $\frac{316.875}{3 \cdot 120} = 0.88...$, for $n = 120$ (number of agents). This demonstrates that even with weights randomly initialized the number of signals per meaning should be quite large.

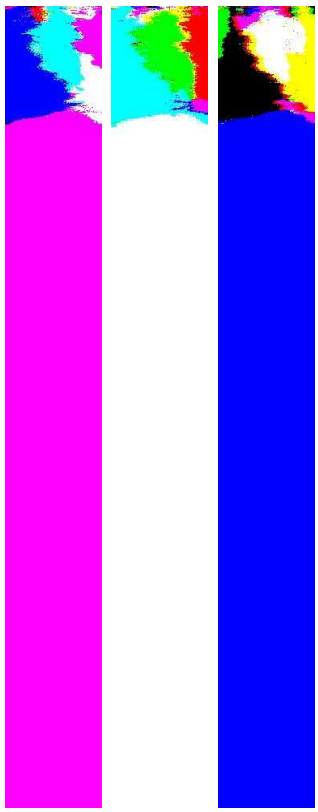

Figure 4.7: $\eta = 0.1$, $\sigma = 2.5$, `inter-child training` allowed. 1000 generations, $t = 80$ communicative episodes for parent-child communication, $t/2 = 40$ for child-child communication. Convergence on a global dialect happens quickly.

of the population of agents when there are only $t = 54$ communicative episodes of child-parent learning and, again, $t/2$ episodes of inter-child learning - hence, a total of 81 communicative episodes, and the situation being thus comparable to the introductory experiment without inter-child learning - it becomes apparent that at least some part of the accelerated convergence was due to the prolonged period of communication among agents, see Figure 4.8. Here, convergence eventuates only after about 400 episodes, a result similar to the result of the earlier experiment.

Why exactly it might seem that inter-child training produces some more homogeneity is an issue with no simple answer. One may not forget that questions are posed here concerning the dynamics of complex and highly sensitive systems, the behavior of which may not always be obvious, or describable by nice formulae; in particular, random differences

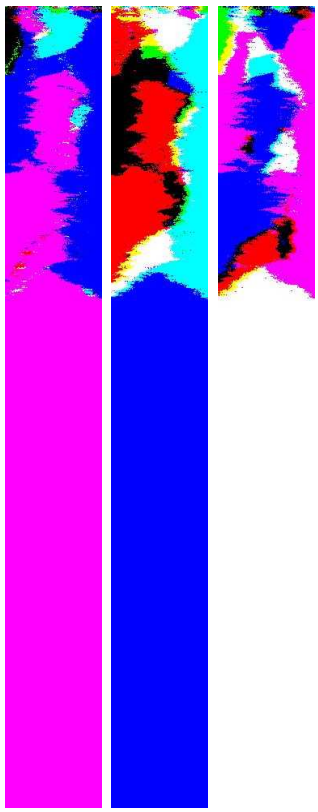

Figure 4.8: $\eta = 0.1$, $\sigma = 2.5$, `inter-child training` allowed. 1000 generations, $t = 54$ communicative episodes for parent-child communication, $t/2 = 27$ for child-child communication. Convergence on a global dialect happens at a later stage, comparable to the situation without inter-child training. As a side note, incidentally, on this map, it has the appearance that the 'blue' dialect on the left-most map is a 'relic' feature (see Chapter 2.2).

in initial conditions may have large effects. I feel[10] that, qualitatively, inter-child training does not result in too different outcomes, and so will in the following, with regard to the additional computational effort required by its implementation, go without it[11].

In fact, another experiment with the same parameter setting (see Figure 4.9) shows that the appearance of additional homogeneity was possibly illusory; there, no convergence on a global dialect occurs during the first 1000 generations, notwithstanding the presence of inter-child training[12].

[10] And here I agree with the opinion expressed in Livingstone (2001).

[11] This also includes experiments with even more complex, or 'realistic', verbal interaction scenarios, including situations where child-parent and parent-parent dialogue are alternating.

[12] A result that could have, and indeed has, been observed when `inter-child training` was not

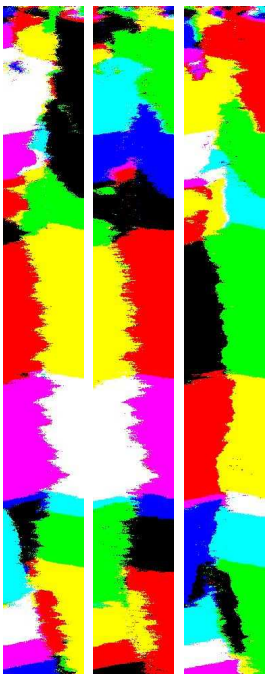

Figure 4.9: Same situation as in Figure 4.8. This time no convergence on a global dialect.

4.4 The role of σ

In order to investigate the role of σ, and, more particularly, its impact on the number of dialects in a population, it is necessary to devise an *effective* method for determining what a dialect actually is, which members of the population belong to a certain language variety, and which to another. Unlike Livingstone and Fyfe (1999), who have measured the hamming distance between two signals for the computation of dialect boundaries[13], I would rather like to concentrate on the traditional criterion for dialects, namely mutual intelligibility (see Chapter 2). However, as will be seen, the two measures are closely related.

The task, still, has not simplified. As has been stated and demonstrated repeatedly in this paper, dialects have no clear boundaries, and isoglosses for different language features take different routes for the same speakers. Therefore, it might occur that there are seemingly

allowed.

[13]Meaning that a dialect was in their work (effectively) defined in terms of common signal use between agents.

few dialects with respect to one meaning agents are negotiating a language system for, whereas for others there are many[14]; at least, there is chance that the intuition about the number of dialects for different meanings among the same population will vary. An appropriate method for dialect determination should, consequently, rely on some sort of averaging over the number of dialects for each language feature available in the community of agents. Such a method will be discussed later.

In the meantime, two other quantities must be investigated; firstly, the correlation between intelligibility and mutual intelligibility, and, secondly, the correlation between speaking and understanding as a measure of dialect outline.

4.4.1 Intelligibility and mutual intelligibility

Intelligibility between two speakers is no symmetric relation. If A understands B, the converse need not be true. Danes are said to comprehend Norwegians without difficulty, whereas the latter do not boast themselves of an as exquisite language ability (Wardhaugh, 2002: p.31). The same (potential) asymmetry holds for the population of artificial agents examined in this paper. However, in both cases - natural and artificial - one would expect that the likelihood of mutual intelligibility increases with the degree of intelligibility displayed by one of the speakers towards the other. Figure 4.10 underscores this intuition.

The diagram displays agent 0's (the leftmost agent in the linear arrangement of the population) intelligibility of other speakers as a function of their position; how good does agent 0 understand agents $1, 2, \ldots, 119$? The second curve depicted shows the "reverse intelligibility", answering the question, "How good do agents $1, 2, \ldots, 119$ comprehend agent 0?" (where the lines agree, only one of them is drawn). It can be clearly seen that most of the time the lines coincide (indeed, in over 70% of the cases), that wherever the lines diverge, they mostly take the same direction; finally, that divergence seems to be slightly more prominent in the latter two thirds of the diagram, that is, it seems to be

[14]This is not what one intuitively expects, however, and it will be seen that dialect numbers for different meanings are indeed closely correlated.

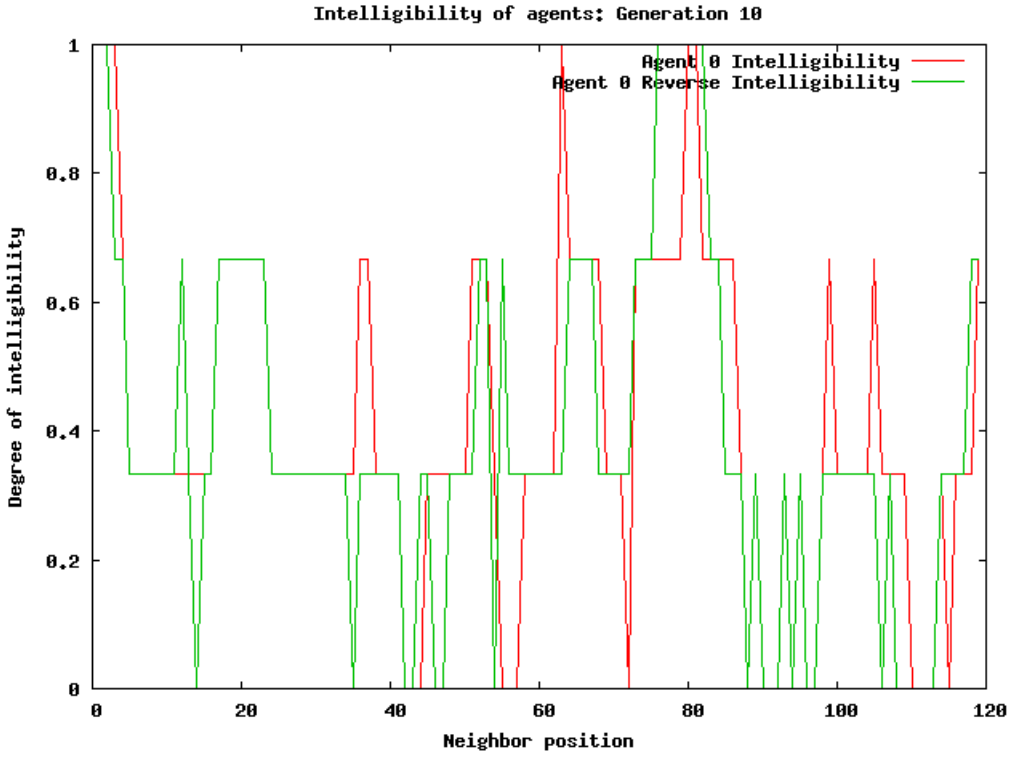

Figure 4.10: $\sigma = 0.5$, $\eta = 0.1$, $t = 80$, 1000 generations. Generation 10, Agent 0. Comparing intelligibility of agent 0 and "reverse intelligibility" in terms of spatial distance to agents of the same population. Both scores take on one of the four values $0, \frac{1}{3}, \frac{2}{3}$, and 1, according to whether one agent understands 0,1,2, or all 3 of the meanings verbalized by the other agent.

slightly increasing with distance. In fact the two measures are equal for 85% of the 1/3 closest neighbors, and for 64% of the two third most distant neighbors, where randomness and chance have greater impact. The correlation coefficient for intelligibility and reverse intelligibility is 0.75 [15], indicating a strong positive relationship between the two measures, and reinforcing the intuitive recognition expressed above.

Repeating this experiment with different agents and generations with the same parameter setting (agents and generations taken from the same civilization) 13 times, a mean correlation coefficient of 0.73 with a variance of 0.02 was obtained. Beyond any doubt is this result significant and entails the pragmatic conclusion that dialect contour may

[15]The *redundancy* value is 0.18, with a complete dependence being signaled by a value of about 0.49. Normalized mutual information is approximately 0.37.

be defined in terms of either 'intelligibility', 'reverse intelligibility', or 'mutual intelligibility' without loss of statistically relevant information. Mutual intelligibility, being the traditional concept, will be chosen in this paper.

4.4.2 Mutual intelligibility and signal use

In order to demonstrate the correlation between mutual intelligibility and signal use[16], a simulation resulting in the dialect maps displayed in Figure 4.11 was conducted[17].

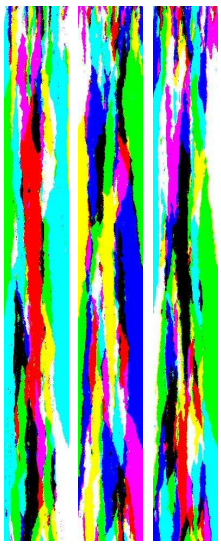

Figure 4.11: $\sigma = 0.5$, $\eta = 0.1$, $t = 80$, 1000 generations.

Arbitrarily choosing generation 500 as a reference point, mutual intelligibility was compared to signal use, or, more appropriately, signal difference[18]. With three meanings existing in the agents' cognitive world and each meaning expressed by a string of length three (and hence, signals used by two agents can have a maximal hamming distance of three), two individual's accumulated signal difference may amount to at most 9. Similarly, the score for mutual intelligibility is taken from the set $\{0, \ldots, 6\}$; an agent may

[16]Note that this analysis is correct only for the settings of the *present* framework. If, for example, the learning rate is set to a smaller or larger value the conclusions need not be true anymore, see next subsection.

[17]The same simulation was the basis for testing the correlation between intelligibility and mutual intelligibility in the previous subsection.

[18]By signal difference I mean the hamming distance between two signals - the number of positions wherein these strings differ - employed by neighboring agents for the same meaning.

understand 0,1, 2, or all 3 meanings another agent is articulating, which is also true for this other agent. Consequently, a *high* score for (this representation of) mutual intelligibility indicates 'language closeness' between two agents, while the same is indicated by a *low* score for signal difference.

Figure 4.12 plots the two measures in one diagram. The discrete points on the 'x axis' represent an agent and its neighbor; in other words, index position, say, 20 on the 'x axis' compares agents 20 and 21.

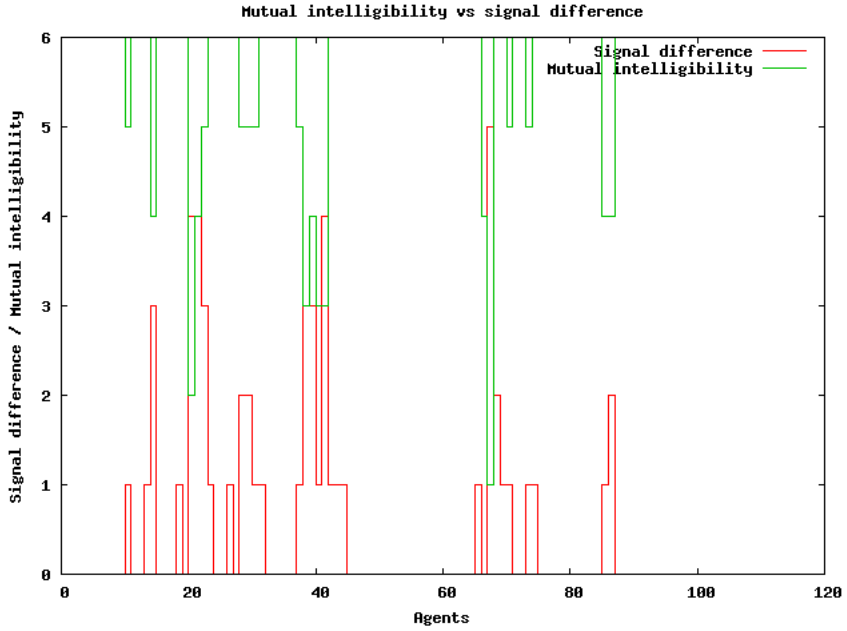

Figure 4.12: $\sigma = 0.5$, $\eta = 0.1$, $t = 80$, 1000 generations. Generation 500 is depicted. The diagram shows difference in signal use (columns at the bottom) and degree of mutual understanding (top) between two immediately neighboring agents. The two measures are negatively correlated.

A negative correlation - as is expected - between signal difference and mutual intelligibility can be discerned. Whenever agents tend to converge in speaking, they also tend to converge in understanding (each other), with the converse true as well. As a matter of fact, in this experiment, whenever mutual intelligibility between two neighboring agents is perfect, then these two agents disagree on at most two bits in the signals used for all three meanings; in 88 out of 100 cases they do not disagree at all. Whenever their

signaling systems are identical (which happens for 89 out of 119 neighbor pairs), then in more than 98% of the cases these agents also have perfect mutual comprehension. The correlation coefficient between mutual intelligibility and signal difference is (-0.33) [19], alluding to the expected negative correlation between the two variables.
Picking randomly 50 other generations from this civilization, I obtained a mean correlation coefficient of (-0.49), with a standard deviation of 0.19. Again, I take this as significant evidence that the two measures are so closely related (for the artificial agents in the experiments conducted in this paper) that using either one of them for the determination of dialect boundaries while neglecting the other does not entail loss of relevant information.

Finally, the general importance of such graphical illustrations as depicted in Figure 4.12 for the determination of the number and shape of dialects in a given generation must be mentioned. One could, of course, use a diagram like this to identify *salient*, or clear-cut, dialect boundaries (as has been suggested, for example, by Livingstone and Fyfe, 1999: p.4). Where signal difference increases, or, sloppy speaking, (statistically) equivalent, mutual intelligibility decreases, a new dialect might be surmised. Were it not for the tediousness of such graphical analysis, this might even turn out to be an adequate solution. The diagram shown in Figure 4.12, for example, promotes the view that there might be four 'main' dialects, the first one including the first 20 agents or so, where a sharp decrease in intelligibility and an accompanying increase in signal difference announces the second main dialect, which includes members up to around agent 40. The third and fourth dialect would then encompass agents 40 to 70 and 70 to 120, respectively[20]. However, two failures of this approach - besides the tediousness - must be emphasized. First, an agent's (speaking and/or listening) behavior might just be an idiosyncratic (or random) behavior with no correlative within the population[21], so that this agent's discrepancy with

[19] The redundancy score is 0.23, with complete dependence signified by a score of 0.42.

[20] Indeed, the method employed for determination of dialect boundaries (see next section) does - more or less - agree with this interpretation.

[21] Although, I admit, this possibility is likely to decrease - at least in the parameter settings investigated so far - with population age, when more systematic language behavior has established.

its two neighbors might not indicate the onset of a new dialect[22]. Second, even with total mutual intelligibility and complete absence of signal difference depicted in the diagram, one cannot conclude that there is just one global language prevalent in the population of agents. A chain of dialects might exist, where complete successive understanding of agents need not indicate that such a relation still holds between the 'leftmost' and the 'rightmost' speaker.

4.4.3 An effective method for dialect determination

The algorithm I devised for the determination of the number of dialects within a given generation's population is as follows. After a generation has performed its communicative duty (i.e. after it has been trained), one of the population's agents is prompted for mutual intelligibility with all other agents. If the resulting value between this former agent and one of the others exceeds a certain threshold[23], both agents are classified as members of the same dialect; since membership in multiple dialect communities for an agent is not desired[24], the latter agent is then removed from the population, and a preliminary dialect community consisting of these two agents is established. Any other agent must now have a sufficient degree of mutual intelligibility with these *two* agents, if its language system is to be classified as belonging to the newly created dialect[25]. This process continues until each member of the generation's population is assigned to exactly one dialect. The - possibly disputable - assumptions of this method are that 1), every agent's language belongs to *exactly* one dialect, and 2) if an agent's language could be assigned to different dialects, then the decision about its affiliation depends on which of these dialects is 'first', that is, which member of one of these dialects is first interrogating for agents with cognate language behavior[26]. Third, dialect determination depends only on intelligibility, not on

[22]In which case, however, the agent's idiolect might be considered a separate dialect.

[23]By default, mutual intelligibility must be attained for at least 2/3 of the utterances expressed by the two agents.

[24]In my implementation, in order to keep the algorithm simple. One *could* of course allow it.

[25]The declarative formulation of the algorithm is thus: a dialect is a set of agents where for all agents i and j the mutual intelligibility score exceeds a certain threshold.

[26]Since order is crucial, agents are tested for mutual intelligibility *in the same order* that they are arranged in their one-dimensional world. Thus, agent i is first testing if its language behavior is close to agent $i+1$, $i+2$, ... with which it had systematic language exchange, instead of first trying to

signal use or other criteria. Finally, the shape and dimension of the resulting dialects is certainly dependent upon the threshold chosen. If this value is too large, that is, a high degree of mutual intelligibility is demanded for dialect membership, it is expected that many dialects will (be supposed to) exist, with few members. If the value is chosen too small, the risk is run of proclaiming a single global dialect; a dialect, though, in which comprehension among (some) speakers may be little more than random guessing.

4.4.4 Experiments with different values of σ

I conducted experiments with nine different values of σ, namely $\sigma = 0.5, 1, 1.5, 2, 2.5, 3, 5, 10,$ and 50 (see Figure 4.13). Before analysing these, some formal definitions are necessary.

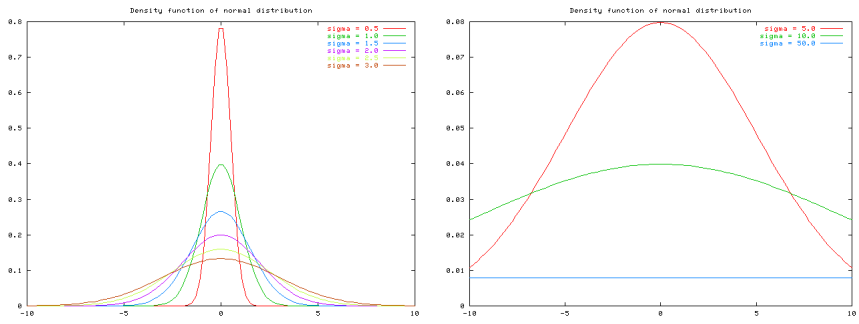

Figure 4.13: Density function of normal distribution for the nine values $\sigma = 0.5, 1, 1.5, 2, 2.5, 3$ (left) and $\sigma = 5, 10, 50$ (right). For $\sigma = 50$ the function looks like a uniform distribution, on this scale.

- Let M denote the set of meanings agents are supposed to verbalize, S the set of possible signals.

- Denote by f_i the meaning-signal mapping performed by agent i, that is, $f_i : M \to S$. Denote by f_i^{-1} the signal-meaning mapping performed by that agent, $f_i^{-1} : S \to M$.[27]

construe a dialect consisting of two very remote (and possibly only accidentally related) speakers, with low probability a finding yet another agent that might accidentally be related to the two.

[27]Note that both functions are well defined and that f_i^{-1}, although the suggestive notation is deliberatively chosen, is not the inverse function of f_i (f_i need not be injective at all).

- *mi_score*,

  $$mi\_score(\text{agent}_i, \text{agent}_j) := \left|\{m \in M \mid f_j^{-1}(f_i(m)) = m\}\right| + \left|\{m \in M \mid f_i^{-1}(f_j(m)) = m\}\right|,$$

  the "mutual intelligibility score".[28]

- *dialect_score*,

  $$dialect\_score(\text{dialect}_i, \text{dialect}_j) := \Big(\sum_{k=1}^{\kappa}\sum_{l=1}^{\lambda} mi\_score(\text{agent}_k^{d_1}, \text{agent}_l^{d_2})\Big)/(\kappa \times \lambda),$$

  where dialect i has κ members $(\text{agent}_1^{d_1},\ldots,\text{agent}_\kappa^{d_1})$ and dialect j has λ. *dialect_score* measures the amount of mutual intelligibility between two dialects.

- Let $D = \{d_1, \ldots, d_n\}$ denote a set of dialects, $n := |D|$. Define

  $$avg\_dialect\_score := \begin{cases} 0 & \text{if } n = 1 \\ \Big(\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} dialect\_score(d_i, d_j)\Big) \times \frac{1}{n(n-1)} & \text{else} \end{cases},$$

  the average dialect score for a set of dialects[29]. It addresses the question, "How well do members of *different* dialects understand each other, on average?"

- $$avg\_identity\_dialect\_score := \frac{\sum_{i=1}^{n} dialect\_score(d_i, d_i)}{n}$$

  answers the question, "How well do members of the *same* dialect understand each other, on average?"

- Let d_i be a dialect, $n_i := |d_i|$ and let $pos(\text{agent}_j)$ denote the position of agent_j in

[28] One could also define a "fractional score" as $\frac{mi\_score}{2|M|}$.

[29] Of course, *dialect_score* is a symmetric relation so that, for computational purposes, the formula can be simplified.

the 1-d world and define[30]

$$dist\_members(d_i) := \sum_{k=1}^{n_i-1} \left| pos(\text{agent}_k^{d_i}) - pos(\text{agent}_{k+1}^{d_i}) \right|,$$

"spatial homogeneity of a dialect". This answers the question, "How much are the members of a dialect dispersed?" with a minimum value of 1, and further define

$$avg\_dist\_members(D) := \frac{\sum_{d \in D} dist\_members(d)}{n},$$

"average spatial homogeneity of a set of dialects."

In order to receive an impression of the impact of σ, Figure 4.14 displays the dialect maps obtained for the nine values of σ (since the signal behavior for a certain value of σ for one meaning is, qualitatively, "just as good" as for any other meaning, only one map for each value of σ is pictured).

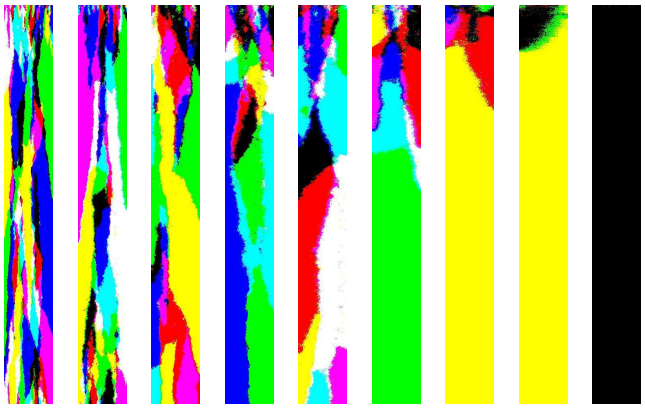

Figure 4.14: Dialect maps (for the meaning 010) for $\sigma = 0.5, 1, 1.5, 2, 2.5, 3, 5, 10$, and 50 (left to right).

How would one expect the variables *avg_dialect_score*, *avg_identity_dialect_score*, and *avg_dist_members* to behave with varying values of σ? What should be expected of the number of dialects? Clearly, the number of dialects should be inversely related to the

[30]Where dialect d_i must be sorted, that is, the members of d_i are enumerated such that their position in the 1-d world increases or decreases.

'size' of σ. On the other hand, if σ increases, *two* possibilities are imaginable, first, the number of dialects decreases, *or*, second, the number of dialects remains the same but mutual intelligibility between different dialects increases (that is, *avg_dialect_score* increases), because speakers are then communicating with more distant listeners, and even if comprehension between two speakers is not adequate to allow the classification of their idiolects as belonging to the same dialect, intelligibility should still be likely to increase. *avg_dist_members* should decrease with σ, since, for example, in the extreme case where there is one global language, the variable will take on its minimum value; such a relationship between the two variables is not compulsory, however, because if, for instance, 'first' and 'last' speaker are constituting one dialect and the rest of the speakers the other, *avg_dist_members* will still be moderately large - and possibly larger than for other scenarios with more dialects (resulting from a *smaller* 'size' of σ). *avg_identity_dialect_score* should be fairly high most of the time (at least as high as the threshold), with a slightly increasing tendency for larger values of σ.

After varying the size of σ, I randomly chose ten generations from the first hundred, the second hundred, etc., generations[31] for each parameter setting and computed the average value for all of the named random variables. Figure 4.15 shows some of the results.

The diagrams confirm (some of) the expectations. The average number of dialects clearly decreases with increasing σ - with a total average of about 12 to 14 for $\sigma = 0.5$ to about 3 to 7 for $\sigma = 2.5$ (within the first 1000 generations). *avg_identity_dialect_score* remains constantly high in all three diagrams (in fact, in all the parameter settings), always well above the default value of 4, in the interval between 5 and 6. *avg_dist_members* *does* decrease with a decreasing number of dialects in the case of $\sigma = 2.0$; in the other two diagrams it performs some zigzag movement, and so does the development of the number of dialects. *avg_dialect_score* - which specifies the mutual understanding between different dialects - *does* seem to slightly increase with σ; the variable remains at an (approximately) constant value of 2 for $\sigma = 0.5$, is slightly larger than 2 for $\sigma = 2.0$ and

[31] Altogether, it were, as before, 1000 generations for each value of σ.

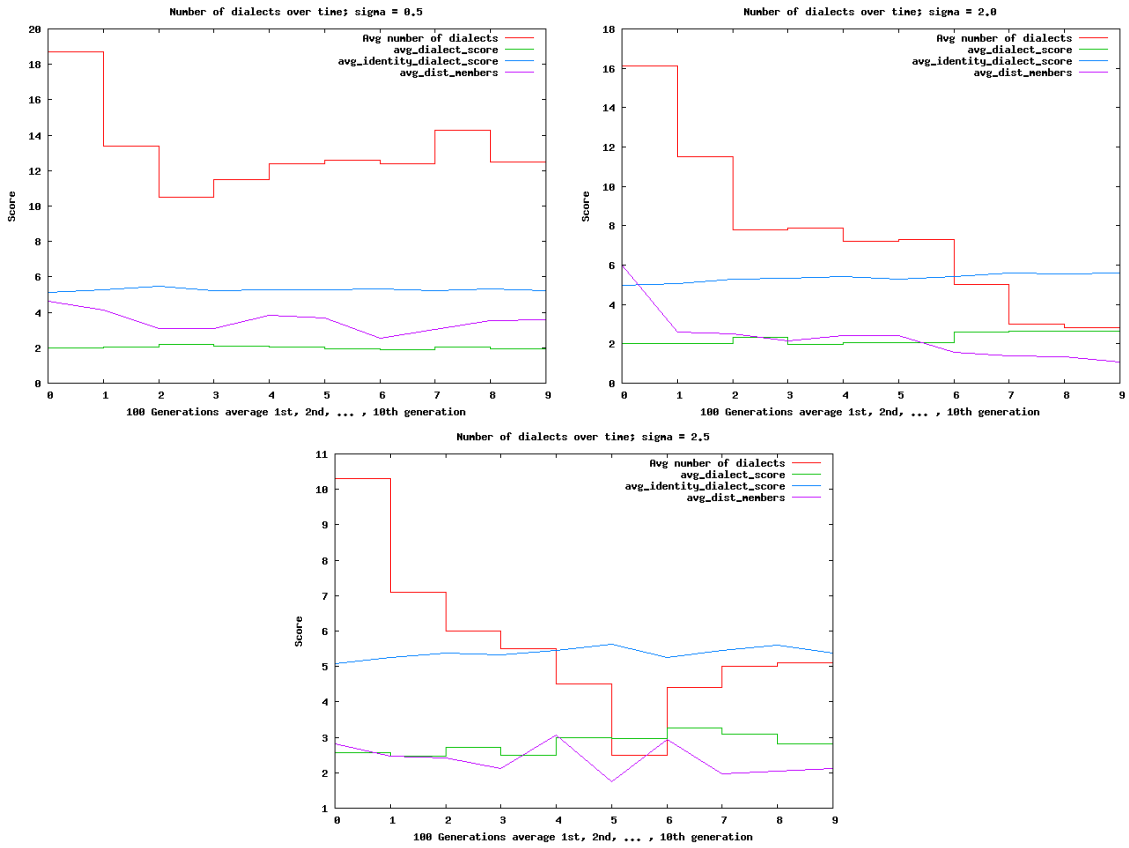

Figure 4.15: Behavior of average number of dialects, *avg_dialect_score*, *avg_identity_dialect_score*, and *avg_dist_members* for $\sigma = 0.5$, $\sigma = 2.0$, and $\sigma = 2.5$ (left to right, top to bottom). Values on the abscissa are 100 generation averages.

in the interval from 2.5 to 3.5 for $\sigma = 2.5$. This proportional relationship seems to be a tendency rather than a law, however, and is not confirmed for all of the experiments; increase of neighborhood size entails - definitely and primarily - reduction of the number of dialects. Only subordinated is the trend to keep dialects distinct but increase mutual intelligibility between them.

σ / score	0.5	1.0	1.5	2.0	2.5	3.0	5.0	10.0	50.0
mean	2.02	2.36	2.55	2.26	2.82	1.28	0.83	0.37	0.03
std	0.09	0.23	0.30	0.28	0.26	1.37	1.37	1.03	0.11

Table 4.1: Mean and standard deviation (of $10 \times 10 = 100$ randomly chosen generations, see above) for *avg_dialect_score* for different values of σ.

Still, the effect is palpable, as demonstrated by Table 4.1; increasing σ also entails - more often than not - increasing mutual intelligibility between different dialects[32]. When σ is too large (here, $\sigma \geq 3.0$) a single global dialect is the likely outcome for later generations, in which case, by its definition (since it is supposed to measure comprehension between *different* dialects), the random variable takes on the value zero.

A plot of the dialect number development for different values of σ in a single diagram will conclude this discussion (see Figure 4.16). No surprises are depicted. Of course, random variation exerts its influence so that, in particular, for small differences of σ results may not always be unambiguous. When the σ's are further apart (as in the second diagram of Figure 4.16), the message becomes more pronounced. There is another remarkable pattern observable in the diagrams. The number of dialects as a function of generation age in a given civilization is at first steadily decreasing, with this decrease coming to a halt after about 200 to 300 generations when either a consolidation can be discerned (if σ is too small for the distinct dialects to converge[33]), or, else, a reduced speed of dialect number decrease that may eventually be destined to result in a global language.

4.5 The role of η

$\eta \in (0, 1)$ controls the speed of learning with which agents adapt to the signaling behavior of other agents (in particular, their parents). Its role in the establishment of shared communicative behavior between agents is to be investigated in the present section. For this reason, experiments have been conducted in which, as before, `inter-child training` was not allowed, agents learned from their parent agents for $t = 80$ episodes, and there were 1000 generations chronicled. σ was set to 3.0, a situation in which convergence towards a global language was observed in the previous section - at least,

[32]With the exception of the behavior, in the table, in the transition from $\sigma = 1.5$ to $\sigma = 2.0$. One may not forget, however, that here the increases of σ are rather small and might be undone by chance and fortune.

[33]The critical value of σ that determines whether individual dialects will (eventually; given enough time) converge (or not), is *suggested* by the diagram in Figure 4.16 to be between $\sigma = 1.5$ and $\sigma = 2.5$.

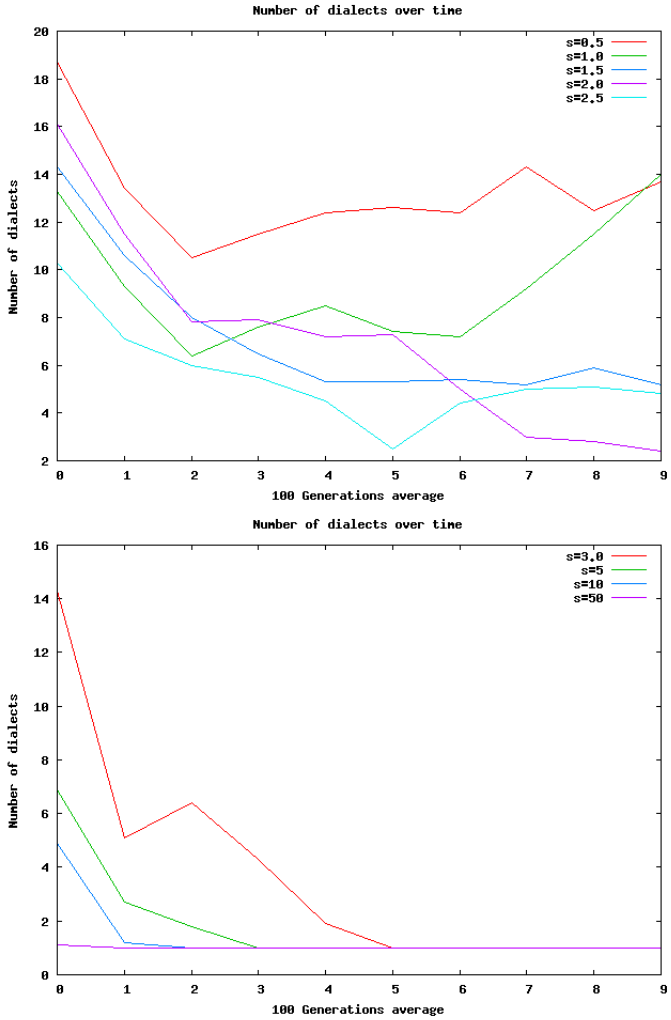

Figure 4.16: Number of dialects as a function of generation 'age'. Top: $\sigma = 0.5$ (topmost line), $\sigma = 1.0, 1.5, 2.0, 2.5$. Bottom: $\sigma = 3.0, 5.0, 10.0, 50.0$ (in this order from top to bottom).

when η was 0.1. The values of η investigated were $\eta = 0.01, 0.02, 0.05, 0.075$, $\eta = 0.1, 0.125, 0.15, 0.2, 0.3, 0.5$, and $\eta = 0.65, 0.8, 0.95$.

4.5.1 When η is (too) small

Suppose η were zero, i.e., no learning would occur at all. Then whatever generation there was, it would not be able to generalize beyond the initial conditions with which it had been equipped at its 'birth'. If agents were randomly initialized, use of a signaling

system would be random likewise. If they were uniformly initialized, all agents would use an *identical* signaling system, although mutual intelligibility were at a low rate of $\frac{1}{n}$, where n denotes the number of meanings agents are equipped with[34].

How does the situation look, then, when η is *close* to zero, but *not* zero?

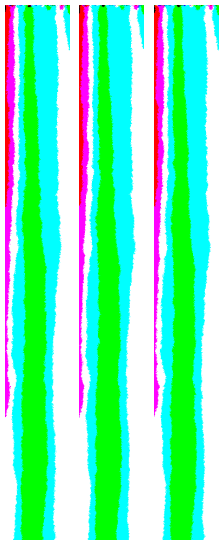

Figure 4.17: $\sigma = 3.0$, $t = 80$, 1000 generations. $\eta = 0.01$. The three maps for the meanings 100, 010, and 001 are (almost) identical.

The fact that the outcome - in this situation – is in many respects (e.g. evolution of a signaling system where all meanings are mapped to the same signal for any given agent) similar to a situation where η is identical to zero turns out to be, despite its alleged 'obviousness', not so easy to explain. I think, there are three phases constituting this development.

Phase 3, once a (child) generation is born all the weights of its members are set to zero (in the framework considered). Suppose that all the members in the parent generation were using a signaling system where for each individual agent each meaning is mapped to the same signal. Then, when a meaning is chosen as a basis for communication between child and parent, the child will have a probability of 1/3 of understanding its parent. If it fails to understand, it will adapt its weights, and it will then *immediately* be able to

[34]Because agents would then map all meanings to the same signal, and so in understanding, the information about the original meaning is lost, and the presented signal mapped consistently to one (and only one) meaning.

understand this signal as the given meaning (since all weights were zero, and have now been moved (if only a little bit) into the 'correct' direction). If, therefore, in the second period of communication, the same meaning is chosen, no learning will occur. If *another* meaning is chosen, the former changes will (either in part or totally) be undone, that is, some or all of the weights will be set to zero again (because, now, the child was taught that a *different* meaning should be associated with the *same* signal). In the course of t communication episodes, this process of 'oscillating around zero' of the weights will continue, with no progress or systematicity. Here, the role of the bias neuron comes into play. This neuron has a consistent desired output of $(+1)$ for all signals presented to the network so that the connections of its weights with the signal neurons can (and indeed, will) be adapted in such a way as to produce the signal hoped for[35]. Each new (signal) example presented to the signal neurons of the learner's network will move these 'bias' weights in the same direction (since, by assumption, all signals presented are identical) so that strong reinforcement is provided for a learner for imitating the 'uniform' signaling behavior of its parent teacher[36]. *Phase 2*, when the communicative behavior of the members of a child's parents is unsystematic (or random), but say, slightly 'favored' for *some* order (mapping from input to output)[37], the operations of the bias neuron might counteract the tendency towards increased signal distinctiveness discussed in Chapter 4.2[38]; since the only genuine consistency is, again, the desired value of the bias neuron and because the weights associated with this neuron will by assumption be more likely to be moved in one direction rather than another, the bias neuron will favor the mapping of *all* meanings to *one* signal, namely to the preferred one amongst the parents. Consequently, the bias neuron might not only transform total (see phase 3) but also partial randomness into regularity. *Phase 1*, I hold that the small size of η is responsible for this unsystemacity, or rather, is accountable for *not* generating systematicity; because

[35]This means that while the other weights oscillate around zero (see above) because they receive contradictory information, it is *solely* the bias neuron that is responsible for generating a systematic output for a given meaning; again, this is because the 'instructions' it receives are consistent.

[36]I.e. to map all meanings to one signal.

[37]As an example, assume that the parent is mapping two meanings to one signal, and the third to another one.

[38]The outcome of the learned behavior will depend on which of the two forces is stronger.

the learning rate is too small, the language of the randomly initialized first generation will retain its random character[39]. Putting this argument in causal order, it can be summarized that if η is too small systematic language behavior will not arise in the first generation(s); this randomness will then be eventually transformed into order by the workings of the bias neuron, yet with the cost of mutual intelligibility. Finally, the fact that there is still change in the system (as opposed to complete regularity), i.e., movement of signal boundaries, is of course due to the existence of neighborhoods in communication.

As an exemplification of the lack of power of a (too) small value of η to introduce fruitful communicative behavior (in terms of intelligibility) into a generation of agents, consider Figure 4.18.

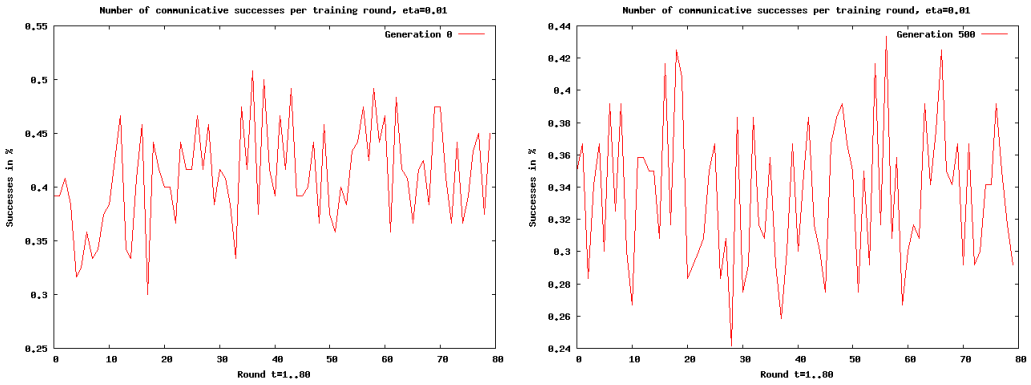

Figure 4.18: $\sigma = 3.0$, $t = 80$, 1000 generations. $\eta = 0.01$. Number of communicative successes between agents as a function of communicative episodes. Left: Generation 0, Right: Generation 500.

It is notable that intelligibility among agents does hardly improve with communicative episodes (as opposed to the observation made when η was ten times larger (see Chapter 4.2)). Also, agents do not evolve better language behavior over time; on the contrary, (even) after 500 generations, understanding amounts to little more than guessing (where guessing would achieve a score of 1/3).

What is, *ceteris paribus*, the minimum value of η necessary for successful language

[39]Additionally, for all subsequent uniformly initialized generations the small value of η will prevent the dissociation of signals for different meanings.

interaction to establish? Experiments with $\eta = 0.02$, that is, a doubling of the factor relative to the analyzed setting, did not show significant improvements. Only when setting it to 0.05, results similar to those observed in the previous sections are found.

4.5.2 When η is (too) large

When η is too close to unity, agents will not be able to converge on (local or global) lects. This is because then whatever is learned from one agent runs the risk of being - completely and immediately - undone by the signal-meaning pair conveyed by another teacher. To be more precise, in this situation the weights of the agents' networks vividly move around in their weight space, settling in (local or global) optima only perchance. Figure 4.19 displays the agents' signaling behavior when η is 0.65, 0.8, and 0.95 (only the signals for one meaning are displayed because the others are quite similar). As the figure suggests, language behavior is still 'less random' when $\eta = 0.65$; no significant differences can be observed between $\eta = 0.8$ and $\eta = 0.95$. The average number of dialects, for example, in the first case is around 40 to 50. In the latter two cases it is constantly 120 (put differently, each agent speaks its own idiolect, without comprehension of *any* of its neighbors other than guessing).

4.5.3 Moderate values of η

When η is (approximately) in the interval $(0.05, 0.6)$, then, in the given setting, a 'normal' dialect development will evolve; i.e. one which results in, for example, populations of agents capable of mutually understanding each other. Figure 4.20 displays the dialect maps effected by such experiments.

Based on the implications suggested by these graphics, a few analyses shall be conducted. A first noticeable impact exercised by an increasing value of η seems to be a 'blurring' of the dialect maps, or, introduction of inhomogeneity. This means when η becomes larger, it seems that neighboring agents are more unlikely to use the same signaling scheme (hence, the hamming distance between their respective signals becomes larger). This impression is corroborated by Figure 4.21. It compares the degree of inhomogeneity

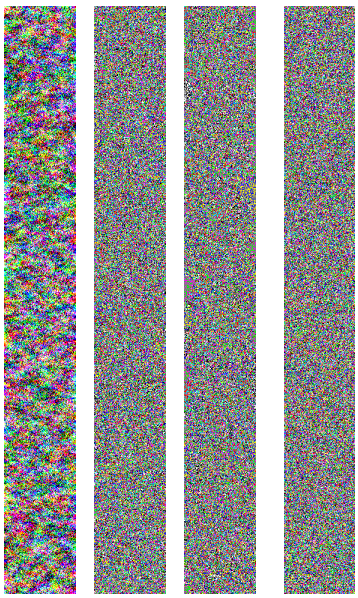

Figure 4.19: $\sigma = 3.0$, $t = 80$, 1000 generations. $\eta = 0.65$, 0.8, and 0.95 (from left to right). Displayed are the signals for the meaning 001. Far right: Random arrangement of dialects.

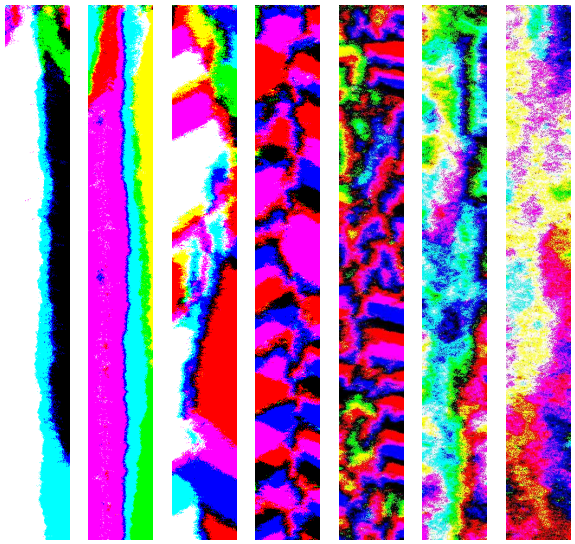

Figure 4.20: $\sigma = 3.0$, $t = 80$, 1000 generations. Values of $\eta = 0.05$, 0.075, 0.125, 0.15, 0.2, 0.3, and 0.5 (from left to right). Depicted are the signal maps for the meaning 001.

(computed by counting the number of different signals, on average, used by two neigh-

boring agents, divided by a normalizing term indicating complete inhomogeneity[40]) for four values of η.

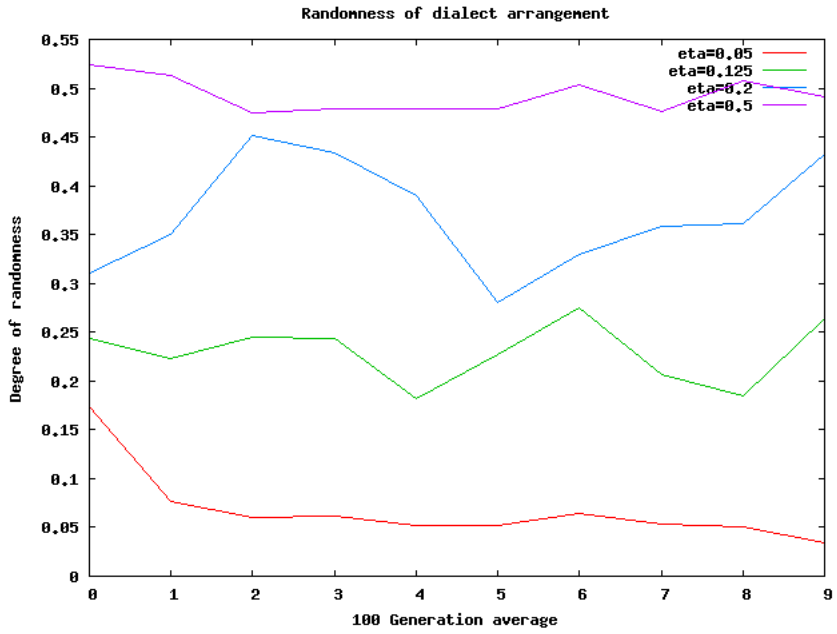

Figure 4.21: The degree of inhomogeneity of a dialect distribution is proportional, on average, to an increase of the learning rate η. The diagram displays the lines for values of $\eta = 0.05, 0.125, 0.2$, and 0.5 (bottom to top).

Interestingly, this tendency towards 'blurring', or inhomogeneity, or 'dispersing' of dialectal behavior does *not* affect (at least, as long as η is moderate) mutual intelligibility of agents within a dialect or between different dialects. On the contrary, agents do, on average, better comprehend each other. Also, evolution of shared language is not slowed down, but accelerated (see Figures 4.22,4.23). That the average spatial displacement of members of one dialect is a function increasing with η is not surprising, on the other hand; the 'blurring' is its manifestation.

How can such an outcome be explained? Speaking metaphorically, if individuals dis-

[40]With the notation introduced in Chapter 4.4.4, the degree of inhomogeneity is calculated as

$$\frac{\frac{|\{f_i(m) \mid f_i(m) \neq f_{i+1}(m),\, m \in M,\, i=1,\ldots,119\}|}{|M|}}{\delta},$$

where $\delta := 120 \cdot \frac{7}{8} = 105$, which is the expected value of the sum of 120 independent random variables X_i, where $X_i = 1$ when $f_i(m) \neq f_{i+1}(m)$ and $X_i = 0$ otherwise. Then $P[X_i = 1] = 7/8$.

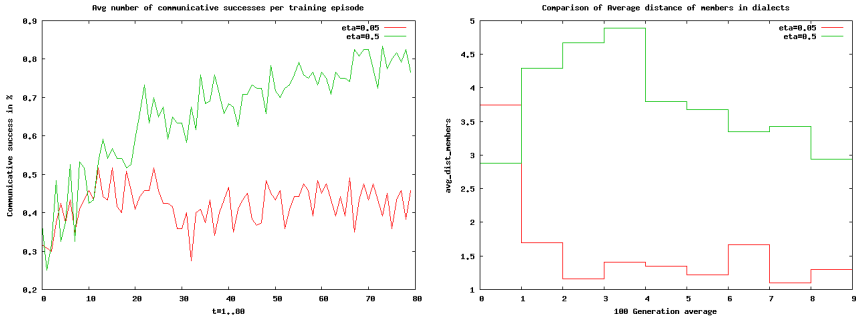

Figure 4.22: Left: Communicative success of agents of the first generation for $\eta = 0.5$ (top line) and $\eta = 0.05$. With a faster learning rate agents are also much faster to establish (locally) a common linguistic system. Right: Average distance of members in the same dialect for $\eta = 0.5$ (top) and $\eta = 0.05$. Whereas for $\eta = 0.05$ agents of the same dialect are spatially 'close' to one another, with a larger η average distance between such individuals is 3 to 4 times larger.

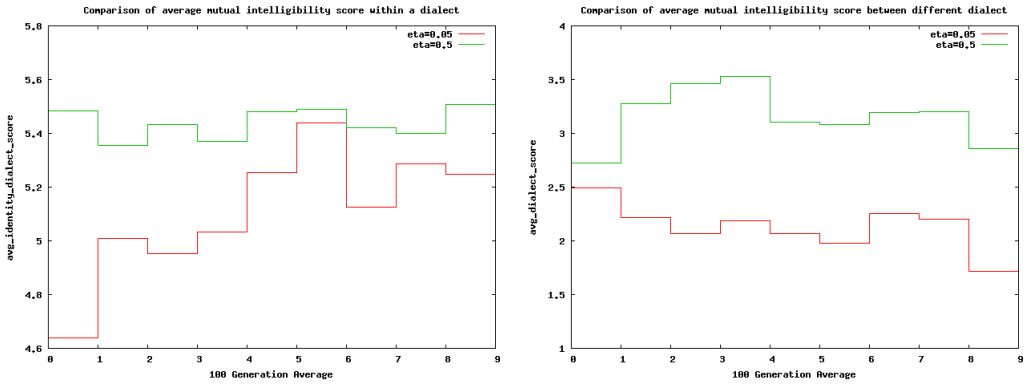

Figure 4.23: How does the learning rate affect mutual intelligibility? Both for understanding within a dialect (left), and for understanding between different dialects, a reasonable increase of η guarantees better scores.

play more 'openness' towards innovation (cf. Wardhaugh, 2002) the likelihood increases that they will adapt to more 'distant' (and hence, 'new') language behavior, whereas, if they are rather resistant, a greater tendency of preservation of the traditional varieties spoken in one's environment can be expected. Transferring this to the artificial agents analyzed in the present paper, one can reason that a larger (again, as long as it remains moderate) value of η will promote the chance of (even) acquiring the language behavior of agents with which communication is occasional, while with a small value of η such

casual interaction will be outweighed by the (likewise only 'resistantly' digested but more frequent) communication with one's closer neighbors. Signaling behavior will thus be more 'dispersed' because agents will be quicker to learn from other agents[41]. With this strong influence exerted by all agents in an agent's local neighborhood, an increase in mutual intelligibility, in general, can concur with this more inhomogeneous use of signals.

η / score	0.05	0.075	0.125	0.15	0.20	0.30	0.50
mean	4.85	5.05	6.35	5.25	8.17	8.24	6.81

Table 4.2: Mean of number of dialects of $10 \times 10 = 100$ randomly chosen generations for each value of η.

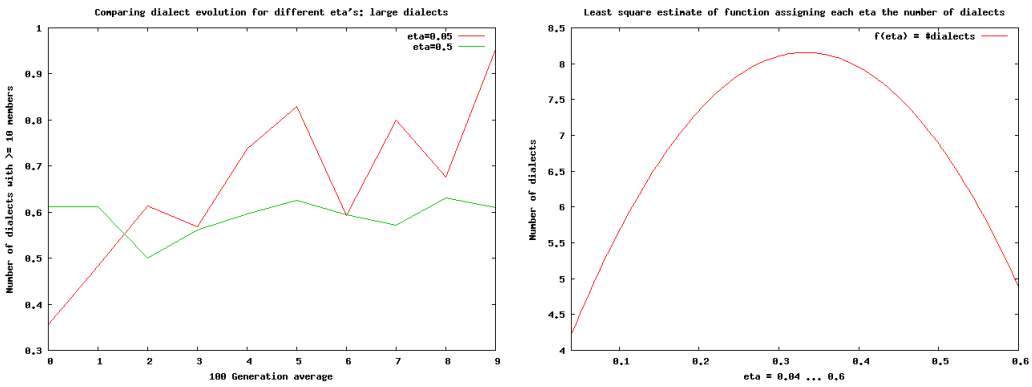

Figure 4.24: Left: Fraction of 'reasonable-sized' dialects (of total number of dialects) for $\eta = 0.05$ (rising line) and $\eta = 0.5$. Right: 2^{nd} order least square interpolation of the data presented in Table 4.2.

To conclude, Figure 4.24 - on the number of dialects as a function of η - exemplifies (again) first, the tendency towards increased homogeneity when η is small. The left diagram plots the fraction of 'reasonable-sized' dialects[42] against generation 'age' for

[41] An instantiation of this logic would be the following example. Suppose an agent quickly adapts its language to its (very close) neighborhood. Suppose further that during its last communicative episode in its lifetime it will encounter a 'foreigner'. If the first agent's learning rate is low, the foreigner will not affect it. If it is moderate, the foreigner might sufficiently change the agent's weight vector to cause (slightly) different signaling performance *without* severely damaging mutual intelligibility with the former's neighbors. If the learning rate is too large, exactly this is what will occur.

[42] Those that have at least 10 members. Many 'small' dialects are an indication of either randomness in the system or local (as opposed to global) organization of language behavior. 'Large' dialects are an indicator of homogeneity.

two values of η. When η is small (increasing line), most dialects initially comprise few members, but gradually homogeneity spreads, with finally more than 90% of the dialects being 'large'. When η is larger, inhomogeneity is given at the beginning and is preserved throughout the evolutionary process. Second, the diagram on the right pictures the 2^{nd} order least square estimate[43] for the function mapping $(0,1) \ni \eta \mapsto \#(\text{of dialects})$ based on the data shown in Table 4.2. The image is not very informative, however, because there are too few data points to commit to reliable estimates; however, it outlines that the number of dialects as a function of η is driven by two forces: with an increase of η mutual intelligibility among agents rises (because they are learning more quickly), which favors fewer dialects. On the other hand, the increased inhomogeneity associated with this situation is an originator of dialect diversity.

As a summary of the workings of η, it has turned out that, in terms of signal use, the smaller η is - and therefore, the more hesitant to change agents are - the more will language behavior be homogeneous among agents; the larger η is - and thus the more 'progressive' agents' 'attitudes' are - the more will language behavior be inhomogeneous and the more 'innovative' will the individual be. In terms of mutual intelligibility, neither too small nor too large values of η are of any worth. If η is 'reasonable-sized', then a larger value promotes quicker learning with no worse performance in comprehension. Considering the increased amount of inhomogeneity this produces, this may also be held liable for preventing convergence to a global uniform standard[44].

4.6 The role of t

The two questions that will be addressed in this section are, "Is it possible to have genuine linguistic behavior evolve - by means of t - where otherwise such behavior would not have developed (due to, in particular, a too small value of η)?" and, more general, "What is the relationship between t and the degree of mutual intelligibility among agents?"

[43] The coefficients are $a = -45.8562$, $b = 30.5412$, $c = 3.0726$ for the interpolating function $f(x) = ax^2 + bx + c$.

[44] Considering only spoken languages, a global uniform dialect is something very unrealistic for human languages.

Concerning the first question, three experiments were conducted, each with 60 agents[45] and a low learning rate parameter η of 0.02, which precluded the advent of mutual intelligibility when t was 80 (see last section), and instead caused the emergence of a system where agents used identical signaling systems for all meanings. The question is whether this outcome is intrinsic to this (small) size of the learning rate parameter or whether mutual intelligibility is a function of *both* η *and* t, possibly depending (primarily) on their relative sizes. Common sense tells me that the latter supposition should be true; 'dull' entities should be able to achieve with lots of work (or, training) as much as 'smart' entities with little. However, while training agents for 150 and 200 communicative episodes per agent (i.e. more than a doubling of t) did indeed improve 'comprehension performance' to a value of about 65 to 75% (as opposed to below 50% with $t = 80$) and promoted the use of two signals for three meanings (as opposed to only one), this result was still not comparable to the setting of, say, $\eta = 0.05$ and $t = 80$ [46], when the agents developed *full* linguistic capacity[47].
Only when setting t to 500 (I did not try out values between 200 and 500, however) was the desired goal achieved. The maps displaying signal use for the three meanings are depicted below (see Figure 4.25).

Concerning the second question, the relationship between t and the degree of mutual intelligibility, Table 4.3 captures the positive correlation between the two variables.

A function whose slope is large for small values and decreases for larger values is, for example, the logarithm. Assuming, therefore, that the true function mapping from t to degree of mutual intelligibility is a logarithm with unknown coefficient a, the least square approximation for the given data points can be calculated; it is shown in Figure 4.26.

Of course, the actual function f mapping from number of communicative episodes to degree of mutual intelligibility is a function of *both* η and t, $f \equiv f(\eta, t)$. As seen, for

[45] Because of the computational costs involved in the computation, with each experiment taking one to two hours on my computer.

[46] Note that the setting $\eta = 0.02$ and $t = 200$ seemed particularly promising because here an increase of t by the factor $\frac{200}{80} = 2.5$ concurred with a decrease of η by the same factor, $\frac{0.05}{0.02} = 2.5$.

[47] By this I mean distinctiveness in signal use and almost complete intelligibility between neighbors.

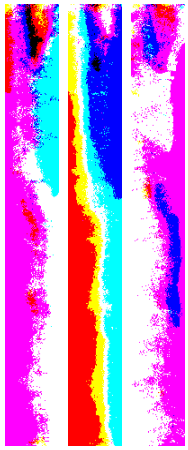

Figure 4.25: $\eta = 0.02$, $\sigma = 3$, $t = 500$, 60 agents.

t	5	7	10	13	16	20	30
mean degree of mutual intelligibility (%)	0.43	0.48	0.58	0.68	0.78	0.94	0.89
average increase	n.d.	0.025	0.033	0.034	0.033	0.038	−0.004

Table 4.3: $\eta = 0.2$, $\sigma = 3.0$, 60 agents, 1000 generations. Different values of t. The mean degree of mutual intelligibility is computed by taking 300 generations at random for each value of t and calculating the average number of communicative successes for these generations. Average increase is computed as the fraction of difference in degree of mutual intelligibility and difference in value of t for two successive t values.

example, when η was 0.02 t needed to be $\gg 200$ in order to obtain a degree of mutual intelligibility comparable to the one achieved for t no larger than 30 when η was 0.2 (see Figure). This means that the coefficient a will be dependent upon η, a_η; it will be smaller for smaller values of η.

Finally, the hypothesis shall be tested that the degree of mutual intelligibility for, say, $t = 7$ is indeed higher than for $t = 5$ [48], in general (for $\eta = 0.2$ fixed), and not only an accident of the random variations exhibited by the particular experiment implemented. Ten experiments for each t are conducted, and the parameters $p \equiv$ degree of mutual

[48] For such small values of t computational complexity is rather low so that hypothesis testing is feasible; with t growing and experiments taking several hours such benefits cannot be afforded.

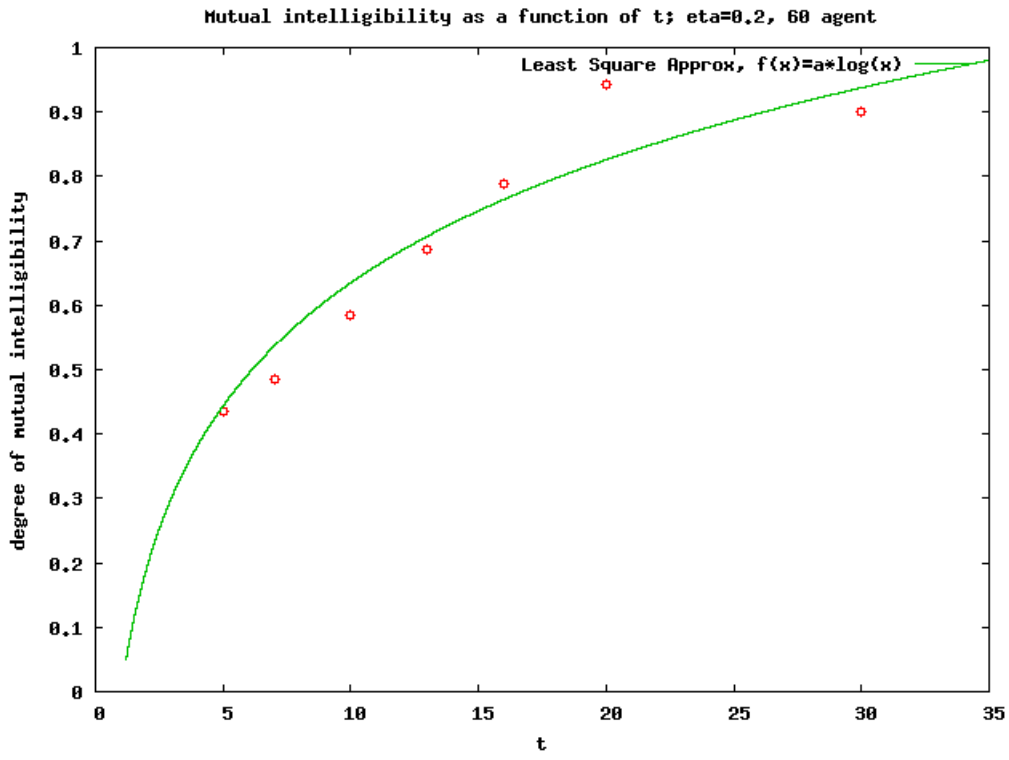

Figure 4.26: $\eta = 0.2$, $\sigma = 3$, 60 agents, least square approximation to data points shown in Table 4.3. Approximation function is supposed to be logarithmic, with a coefficient of $a = 0.27569$.

intelligibility for $t = 5$ and $q \equiv$ degree of mutual intelligibility for $t = 7$ are estimated as

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} p_i = 0.436, \quad \hat{q} = \frac{1}{n}\sum_{i=1}^{n} q_i = 0.484,$$

where $n = 10$ and p_i is the average degree of mutual intelligibility for experiment i for $t = 5$; q_i likewise. The estimated difference $\hat{d}$ is computed as $\hat{d} := \hat{q} - \hat{p} = 0.047$ and its variance is given by (cf. Mitchell, 1995: p.144)

$$\sigma_{\hat{d}}^2 = \frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{q}(1-\hat{q})}{n} = 0.04956$$

From this the probability that the *true* difference between $t = 5$ and $t = 7$ (with respect to degree of mutual intelligibility, and for fixed η), d, is greater 0, given that $\hat{d}$ is 0.047, can be calculated using the normal approximation to the binomial distribution; with

$\sigma_{\hat{d}} = \sqrt{0.04956} = 0.22262$, it is[49]

$$\Pr[\hat{d} < \mu_{\hat{d}} + 0.047] = \Pr[\hat{d} < \mu_{\hat{d}} + 0.21\sigma_{\hat{d}}] = 0.58,$$

which means that the probability that with $t = 5$ the (true) degree of mutual intelligibility is smaller than with $t = 7$ is, given the observed data sample, almost 60%; in other words, the hypothesis that $t = 7$ is increasing the amount of mutual intelligibility among agents over $t = 5$ (with $\eta = 0.2$), is accepted with confidence 0.58 (with a larger sample size, I am positive, confidence would increase because there is clear - intuitive as well as argumentative - rationale that the amount of verbal interaction between agents should enhance their ability to agree on a (locally) shared language system[50]).

4.7 The role of redundancy

It was noted in Chapter 2.4 that language change might be attributable (at least) in part to the amount of redundancy available in a linguistic system. When some structural device is unused but present it might provide the basis for innovation - where otherwise such innovation would not occur because of the indispensability and vitality of the remaining (used and useful) resources.

This is to be investigated in the present section. In the implementations discussed so far, a meaning set of cardinality three had to be negotiated with the help of eight possible signals. A situation where there are less (possible) signals than 'concepts to talk about' is certainly not interesting. Misunderstandings will be the custom. Mutual intelligibility will decrease. Above all, given their dual level of patterning, the scenario seems to be very unlikely, if not altogether impossible, for human languages. Additionally, cases where one (or two) signal(s) were representative of all three meanings were considered earlier (if, then, for other reasons, namely a too low learning rate parameter).

[49] Note that $d > 0 \iff d > \hat{d} - \hat{d} \iff \hat{d} < d + 0.047$ and $\mu_{\hat{d}} := \mathrm{E}[\hat{d}] = d$.

[50] Probably this should be stressed more strongly. I do not claim that there is only 60% confidence that the amount of verbal interaction will positively influence agents' linguistic capacities, and that there *might* be even higher chance but one does not know. What I am saying, instead, is that I am lacking the time to do extensive hypothesis testing and that 60% confidence were readily obtained.

Thus - denoting by m and n the number of (actual) meanings and the number (possible) signals respectively - the focus of the present chapter will be on situations where always $n \geq m$; two particular instances are of interest, namely, $n \gg m$, and, even more so, $n = m$, because they are likely candidates for demonstrating the role of redundancy in the diffusion of variation; if the theory is valid, in the first case change and variation should be ubiquitous whilst in the second they should be considerably reduced, *ceteris paribus*.

Considering the second case, $n = m$, an experiment was arranged where existence of variety was a permanent fact under parameter settings investigated earlier. η was set to 0.2 and the neighborhood size parameter σ to 3.0, a situation where approximately 6 to 7 dialects had emerged, on average, with $n = 8$ and $m = 3$. In order not to renounce on the benefits of graphical analysis and illustration, n was kept constant at $2^3 = 8$. The number of input neurons could thus also be retained unmodified, but this time 'binary meanings' were implemented instead of a winner-take-all competition over 'unary meanings' (see Chapter 3). Prior to the experiment, I was skeptical about the effect of equalizing n and m because I thought that a primary force for change would be the availability of dialects[51], which - due to the geographical arrangement of the agents - would always be present in my experiments. Before turning to the actual results, a very brief illustration of the impact of $m = n$ shall be given (see Figure 4.27).

It can be discerned that the agents are still successful at evolving (locally or globally) common language behavior. For each generation, communicative success is increasing with the amount of training[52]. Later generations are more successful than earlier ones, again a fact similar to what has been observed previously. Yet, agents are not as successful any more as they used to be with $m = 3$. This is a 'natural' development; with more meanings to talk about, *guessing* the right 'topic' will become harder. Also, with redundancy absent[53], agents are unable to develop lexicons with synonymous entries so

[51] Because with the existence of dialects, individuals always have the possibility of imitating (or learning from) 'foreigners', which could initiate change (cf. Wardhaugh, 2002).

[52] That is, language use.

[53] And (since all parameters are in their 'normal' range) the agents' drive for distinctiveness in signal use still present (that is, a drive for exactly one signal per meaning (see Chapter 4.2)).

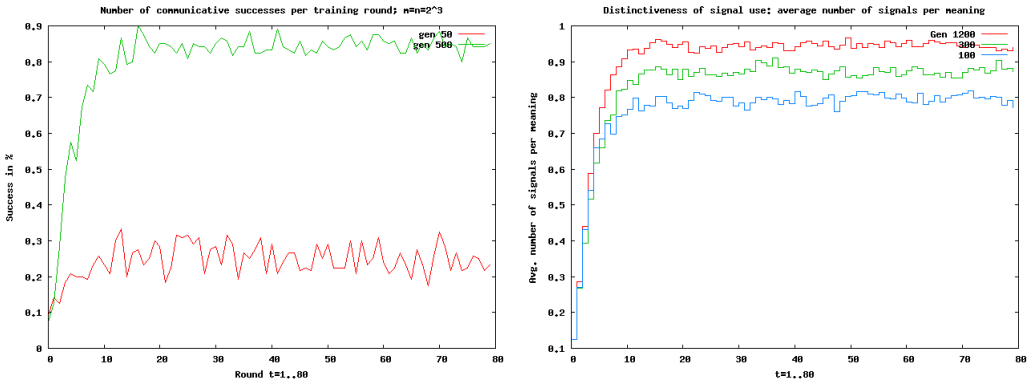

Figure 4.27: $\eta = 0.2$, $\sigma = 3.0$, $t = 80$, 1200 generations. $n = m = 2^3$. Left: with a value of communicative success of around 0.9 for generation 500, agents performs considerably worse than in a situation where redundancy in signal use was available (see Figure 4.4; generation 50 even performs worse than did generation 0 in the earlier experiment). Right: Agents still evolve distinct signals for each meaning, but are likewise less successful than in earlier experiments, never attaining 'perfect' distinctiveness.

that mutual intelligibility will suffer - if only slightly[54]. In general, no significant differences can be recognized however, and, at least for the scenario under analysis here, equating m and n does not (severely) deteriorate or influence agents' abilities to talk to and understand each other. This holds, too, for the number of signals per meaning employed by the average agent (right diagram in Figure 4.27); like in earlier experiments, agents will tend to have their mappings $M \rightarrow S$ injective (compare Figure 4.6).

More astounding than this is the result revealed by the dialect maps. It seems that redundancy is *indeed* one of the major forces driving change and variation. In a parameter setting selected deliberately because of its seeming resistance to convergence to a global dialect when redundancy is available, the agents' languages will *now* - after a chaotic beginning - eventually exhibit zero variation in that all of them speak the same - static and immutable - global language (see Figure 4.28). Owing to the import of this finding, I have repeated the experiment four times; in two cases convergence was observed after several hundred generations. In the others, such dynamics was not observed but, similarly, an eventual predominance of one of the dialects, with the outlook that this dialect

[54]This is true, of course, because if I know that the concept *house* is associated with the forms *house* or *domicile*, I will be able to understand both my educated left and my working-class right neighbor. If I have no such storing capacity, I can at most understand one of them.

was supposed to be the global language if only it had been given enough time (compare Figure 4.29).

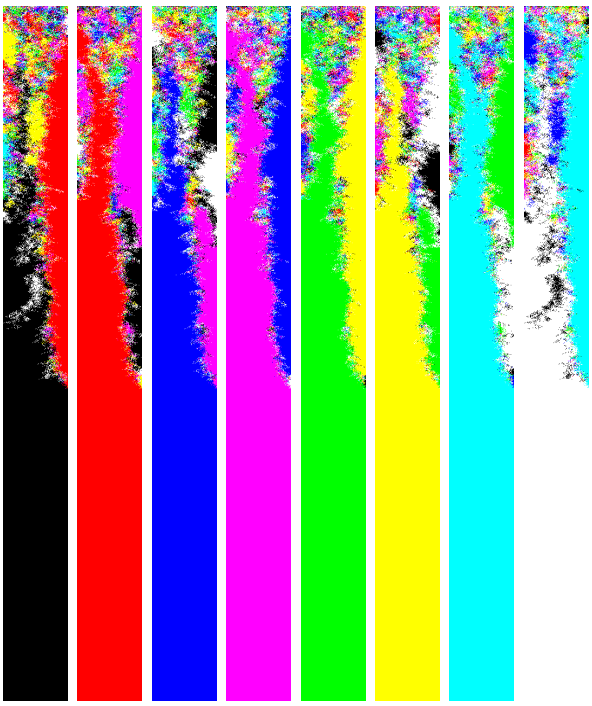

Figure 4.28: $\eta = 0.2$, $\sigma = 3.0$, $t = 80$, 1000 generations. No redundancy. Initial chaotic language behavior is succeeded by convergence to a global, static dialect. This is in contrast to the outcome observed when the signaling system was redundant (compare Figure 4.20). Displayed are the maps for all $m = n = 2^3$ meanings.

I take this as strong empirical support for the theoretical reasoning exercised by Lass (1997) that redundancy (or, junk) might be the *prima causa* of change and variation. Taking the samples considered here as reference, one might hypothesize that under all 'normal' circumstances (parameter settings) change would be inhibited - given enough time - if redundancy was not existent.

The case $n \gg m$ is illustrated in Figure 4.30.

4.8 The role of agent initialization

What happens when all 'newborn' agents are initialized randomly rather than uniformly? Will this be a negligible side effect, or will the so generated 'random diversity' have a

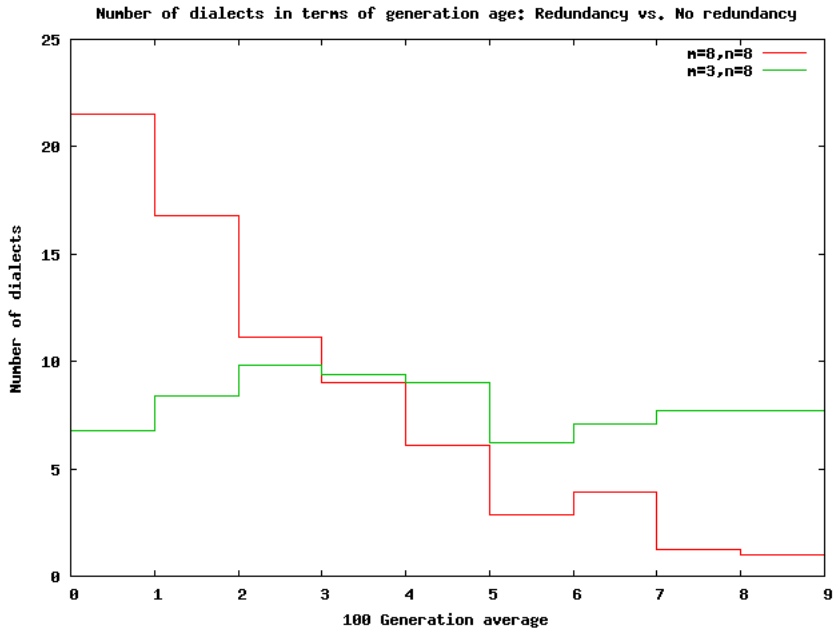

Figure 4.29: $\eta = 0.2$, $\sigma = 3.0$. Number of dialects as a function of generation age. With redundancy in the signaling system, variation is preserved throughout the civilization's lifetime. When redundancy is absent (decreasing line), convergence will occur. The initial increase in the number of dialects for the latter case is due to a decrease, in the beginning, in mutual intelligibility associated with this situation (see text for explanation). The curve depicting the number of dialects without redundancy is an average of four experiments.

chance of being preserved throughout a civilization's evolution? Will it prevent the emergence of a global dialect? Table 4.4 analyses the effect of random agent initialization on communicative success among agents; Figure 4.31 its influence on the number of dialects over a civilization's lifetime.

4.9 The role of noise

It has been noted (e.g., Livingstone and Fyfe, 1999) that perfect replication of signaling schemes is not a property of human language acquisition because of errors in language production or comprehension, on the part of the learner or the teachers. It has also been suggested that such noise is a possible originator of language change (Steels and Kaplan, 1998).

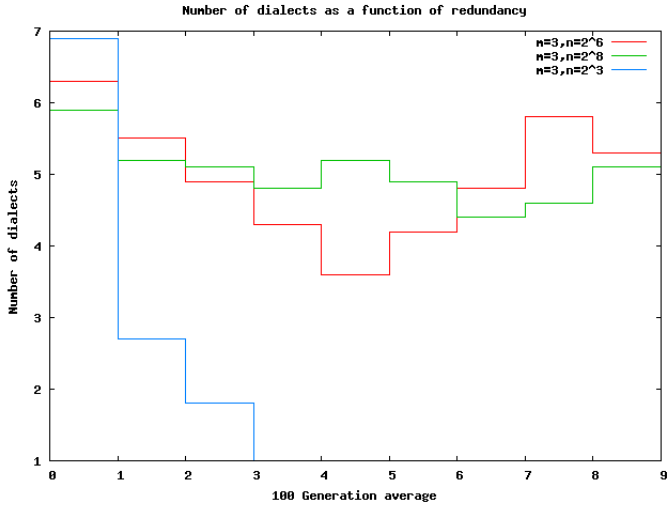

Figure 4.30: $\eta = 0.1$, $\sigma = 5.0$, $n = 2^3, 2^6$, and 2^8, $m = 3$. When $n \gg m$, diversity can be upheld in a situation where it cannot when $n \approxeq m$ (decreasing line). Obviously, when redundancy is large enough, an additional increase will not have much impact - if any - anymore.

σ / comm. succ	3.0	5.0	50.0
random init (%)	79.1	80.3	82.8
uniform init (%)	96.9	98.6	99.9
uniform and high redundancy		97.9	

Table 4.4: $\eta = 0.1$, $t = 80$, 120 agents, 1000 generations, $m = 3$, $n = 2^3$ (2^8 for high redundancy). $\sigma = 3.0$, 5.0, and 50.0 respectively. The average number of communicative success (average over all generations) increases with σ, but random initialization significantly reduces the basic level of success. This is no surprise, of course, because randomness encourages disagreement. The last line in the table suggests that high redundancy does not result in such communication barrier, although it is likewise an originator of diversity, see figure below.

What I wish to investigate in the present section is the question whether noise can sustain variation where it would have subsided under conditions discussed earlier. Modeling noise as the random inversion of bits in language production[55], it turned out that if this chance error was too high[56] it would render language predictability impossible, and thus prevent the emergence of intelligible shared language. With a sufficiently low level of

[55]Inversion in language comprehension would be possible as well, but was omitted here for the sake of simplicity of the model.

[56]Where an error rate of 5% turned out to be too high already.

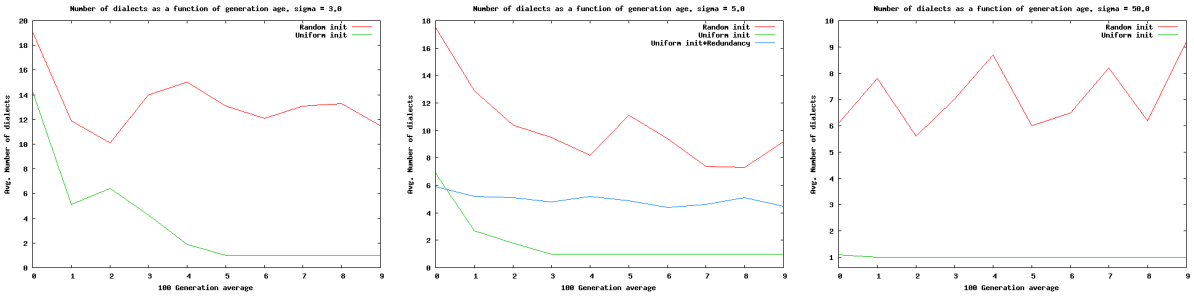

Figure 4.31: $\eta = 0.1$, $t = 80$, 120 agents, 1000 generations, $n = 2^3$, $m = 3$. Number of dialects as a function of generation age. $\sigma = 3.0$, 5.0, and 50.0 (from left to right). Each diagram compares random and uniform initialization (bottom lines in diagrams) of agents. It can be summarized 1) that random initialization preserves diversity where uniform initialization does not, 2) that - independent of the kind of initialization - the number of dialects decreases when neighborhood size increases. Finally, the second diagram illustrates that the effect of random initialization is more severe (in that it entails 'more' diversity) than that of uniform initialization plus large increment of redundancy (blue curve in the middle; compare Figure 4.30).

noise (about 0.5%), however, language comprehension between agents was retained and variation could still be perpetuated, as long as other forces were of 'moderate' strength, in situations where absence of noise implied convergence to a global dialect (see Figure 4.32).

4.10 Summary

To conclude this chapter, a summary of the results obtained will be provided. First it was seen that the dialects generated by the artificial agents employed in this work displayed features similar to those observed in human languages; namely, indistinct dialect boundaries, criss-crossing isoglosses, and, in general, 'change'. Next, the importance of several parameters was investigated, among these, the neighborhood size, `inter-child training`, the learning rate, and the amount of training, t, each agent received upon entering a language community. It could be deduced that the neighborhood size was negatively correlated with the number of dialects, that the success of language learning was independent of whether child agents would learn from parents only or whether they

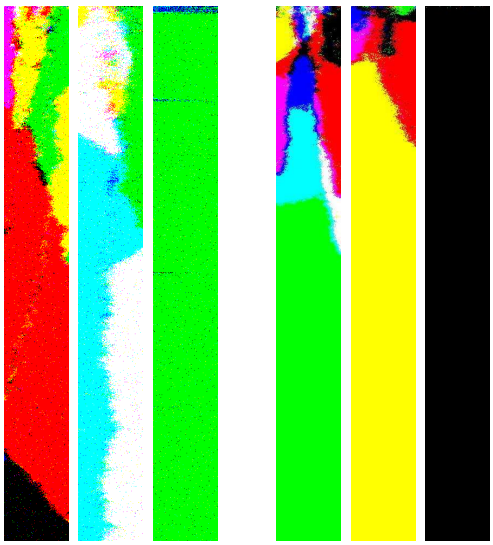

Figure 4.32: $\eta = 0.1$, $t = 80$, 120 agents, 1000 generations, $n = 2^3$, $m = 3$. Left: Signaling behavior of the agents for the meaning 100 for $\sigma = 3.0$, 5.0, and 50.0. On the right: the respective dialect maps when noise was absent, repeated here for the sake of convenience from Figure 4.14.

additionally spoke with their fellows, and instead success was foremost a function of the amount of training they were exposed to. Upon assessing these results, it must be noted that the analysis undertaken in the present chapter was implemented in the name of traditional dialectology, with no reference, for example, to sociolinguistic aspects of language variety. A first step in this direction was taken in the discussion of η. It was shown that - as long as it was 'moderate' - a larger value of η promoted quicker language learning and guaranteed slightly better performance. Thus, η could be interpreted as an agent's status or an agent's 'group index'; when an agent is considered to be of 'high rank' in the community of language users, fellows will communicate with it under the premise of a 'large' learning rate parameter. Likewise, a group's 'connectivity' might be controlled by η, one of the questions investigated in the next chapter.

Finally, issues of language change were tackled. Under which conditions would language change be inhibited, under which conditions would it be rendered feasible? It was seen that the joint absence of noise (imperfect learning), random agent initialization (which could be subsumed under the heading of noise), and redundancy in signal use implied the

absence of language variation in many situations, but that the lack of either one of these parameters could be substituted by the others in the aim of generating change. If one thus wants to be as presumptous as to generalize these observations to human languages, one *could* draw the conclusion that a search for a single hypothesis for the explanation of language change is an unduly simplification of the processes involved and that instead many rivers join to form the sea[57]. A second conclusion that might be drawn is that an appeal to functional explanations for language change does not necessarily need to be made; by their construction, the artificial agents had no explicit assessment of language states, and thus could not explicitly favor one over the other. Still, with (stochastical) parameters set appropriately, change did happen.

[57] As noted earlier, however, such a conclusion need not be drawn. The experiments conducted merely show that neither of the alleged causations of language change nominated in Chapter 2.4 must be implausible. It might, however, well be that in reality, say, redundancy is exclusively responsible for language variation because all of the other factors are - possibly - underrepresented.

Chapter 5

New Simulations

> If, as we argue, much of the structure of language is emergent, then a modeling methodology is appropriate, since it is notoriously difficult to come up with reliable intuitions about emergent behaviors (Kirby and Hurford, 2002: p.143).

5.1 Introduction

Having analyzed some of the parameters included in the model introduced in Chapter 3, the present chapter is designed to investigate some of the framework's other possibilities; what kind of languages can be evolved - if indeed they can - relying solely on the means discussed, that is, a set of agents represented as neural networks and a simple learning rule to modify the strength of the connections of their internal units? So far, the communication systems developed had the layout of lexicons. Agents successfully learned to associate meanings with signals; in Saussurean terms, they learned the usage of a 'sign'. Is it possible that more complicated linguistic structures can be evolved using only the restricted mechanisms constituting this computational model, dispensing with (sophisticated) innate biases and the like? Is language acquisition and evolution 'a miracle' ? Is it a catastrophic accident of human genetic evolution (cf. Pinker and Bloom, 1990) and thus reserved solely for the capacities of human biology? Is language dependent upon a specifically operating language organ, or can general reasoning abilities be held responsible for it?

Before answers concerning some of these very involved questions (for which the present

discussion can, of course, not hope to find definite solutions) can be envisaged, the evolution of linguistic systems more subtle than the ones discussed so far shall be sketched. The next section describes the evolution of compositional signal structure from compositional meaning structure[1]; section 5.3 will then outline the evolution of a *grammar*[2] within the possibilities of the present framework. Finally, section 5.4 will present a computational approach to sociolinguistic factors of language variation.

5.2 Deriving a compositional language

Frege's compositionality principle states that the meaning of a complex sentence is a function of the meaning of its parts and the way these are put together. It shall be tested in the present section whether the languages evolved by the processes described in the last chapters can or do have this property. In a first simple experiment, 8 agents were used, each trained to associate eight different binary meanings with some signal. Noise was not allowed. The emergence of dialectal patterns was not an issue so that a large value of the neighborhood size σ was chosen; it was set to 50. η was chosen from what had earlier been recognized as a plausible value, namely 0.15; t was set to 20 and a total of 300 generations were monitored. The meanings the agents were supposed to verbalize *shall* be given the following natural language interpretation.

	sleep	**eat**
I	00*0*	00**1**
you	01*0*	01**1**
she	10*0*	10**1**
we	11*0*	11**1**

Table 5.1: Possible interpretation of the eight binary symbols as being composed of two verbs (eat, sleep) and four personal pronouns. The binary correspondence of the English pronouns can be read off the table (e.g. you = 01, etc.).

When examining the signaling pattern of the last generation of one particular evo-

[1]Compositionality might have been present in some of the experiments of the last chapter, but was not an issue and hence not examined for.

[2]That is, a (variable) *sequence* of symbols rather than fixed length tokens.

lution taking place under the conditions just outlined, it is found that all agents have developed identical meaning-signal associations. Their lexicons are displayed in the following table[3] [4].

meaning	signal
000	bba
010	baa
100	bbb
110	bab
001	aba
011	aaa
101	abb
111	aab

Table 5.2: Meaning-signal mapping learned by the agents after 300 generations in the setting described in the text. The signaling system is compositional.

It can be seen that the evolved signaling system is compositional. If $Pred \in \{0,1\}$ and $Arg = (Arg_1, Arg_2) \in \{0,1\}^2$ denote the verb predicate and the pronoun argument respectively, then the function mapping from meanings to forms has the logical form $f(Arg \wedge Pred) = S(\neg Pred) + (S(\neg Arg_2) + S(Arg_1))$, where $+$ is to denote symbol concatenation and S is the 'signal assignment' function $S(0) := a$, $S(1) := b$ [5]. In terms of language classification, the agents have hence evolved a VS language. Conversely, the meaning of the complex sentence VP+NP is a result of the meaning of the simple sentence VP and the simple sentence NP, $g(\mathrm{VP} + \mathrm{NP}) = h(M(\mathrm{NP})) \wedge \neg(M(\mathrm{VP}))$, where h is a function with similar properties as illustrated above for the inverse function f, and the 'meaning' assignment function M is defined as $M(a) := 0$, $M(b) := 1$ [6]. Hence, the language is compositional by the above definition.

It should be noted that, first, the exact outcome of the experiment (in terms of, e.g.,

[3] In the following, syntactic symbols are denoted by lower case Latin characters a, b and semantic symbols by the Arabic numbers $0, 1$. This has nothing to do with implementational issues where bits are mapped onto bits, but with clearness of representation.

[4] It should be noted that the lexicons of agents of 'early' generations are 'not yet' compositional or identical. Deriving compositionality is likewise a process that has to be negotiated among agents.

[5] For example, $f(\text{I eat}) = f(00 \wedge 1) = S(0) + (S(1)S(0)) = aba$, see Table 5.2

[6] The notation here is not to be taken rigorously mathematical, but rather as an alluding symbolization. The only thing to be shown is that a sentence can be assigned its meaning by *regular* transformations.

VS or SV order) is dependent upon initial conditions and the dynamics of the described process of shared lexicon evolution - rather than, to add, genetically inherited 'principles and parameters'. Second, it was not obvious in the first place that the derived language should be compositional, because although it was seen earlier that agents would develop bijective mappings from meanings to forms, these could have turned out to be 'hard-wired' and unsystematic (to give an example of an irregular mapping, if the signals for "I sleep" and "you sleep" are exchanged in Table 5.2, with the rest of the mappings unaltered, then there will be no 'compositional' function assigning correct interpretation to all of the signals).

Redundancy and chain shifts

In another experiment where, unlike in the experiment discussed above, redundancy (and hence diversity) was present ($m = 3$ and $n = 4$), it turned out that members of the same dialect likewise evolved compositional signaling systems but tended to disagree on the redundant bit; for example, agents developed syntactic forms where the logical predicate was coded in the last bit of the 'sentence' and the arguments occupied the preceding two bits. The sentence's initial bit carried no meaning and so was instantiated differently by different agents. For example, the sentence $b + \mathbf{aa} + a$ had the meaning $\mathbf{11}0$ (= *sleep*(**we**)) for most of the agents but some also used the sentence $a + aa + a$. After some generations, a 'chain shift' was initiated when the meaning *sleep* was also sometimes instantiated as the symbol b, thus colliding with the form for the meaning *eat*. During this period, where compositionality was partly restrained, the "superfluous" initial bit took over the function of differentiation. This process was only complete when *sleep* had 'pushed' *eat* further, to be (consistently) realized as a (while *sleep* was then realized as b). It was then that full compositionality had returned to the system and the initial bit had regained its 'junk' function, possibly designed to be the medium of yet another language change.

Changes to the Meaning Space

What happens when agents are not trained on a certain meaning, as when one of the meanings representable by three binary digits is omitted from the training 'corpus'? Will the signaling system be 'compositional enough' to reproduce this meaning even though no agent has ever encountered it? In a test, it turned out that this is the case: when agents were prompted to verbalize the neglected meaning and afterwards to interpret their own signal, they were able to retrieve the never seen meaning in almost 70% of the cases, which is a more than 5 times higher score than guessing.

What happens when some meanings are prompted more frequently for verbalization than others? As has been outlined by Kirby and Hurford (2002), the top ten verbs of English ranked by frequency are irregular. The idea is that high-frequency concepts are harder to adapt to a regular pattern of language use, simply because of their commonality. When therefore implementing a 'Zipfian frequency distribution', as they call it, over their meaning space, they actually obtain the result of irregularity, or non-compositionality, for these more commonly verbalized items. In my own simulations, it has turned out that irregularity could not be generated this way. I think this is due to the fact that their model is based upon successive generalizations of context-free grammar rules, which are inhibited by the frequent encountering of tokens, and irregularity ensues. In my model, on the other hand, compositionality seems to stem simply from the layout of the neural network with its tendency to map similar inputs to similar outputs (because this similar input is mediated by the *same* weights). If some meanings are drawn more frequently, they will be responsible for the initial size of the weights, which the other (less frequent) meanings will have to accept as a fixed *datum* - in other words, while in the case of a uniform distribution over the meaning space compositionality is the joint effort of the co-evolution of different meaning-form associations, it will be dictated by the more frequent tokens in the case of a 'Zipfian distribution'.
The distinctness of the models employed leads thus to the difference in the results observed.

5.3 Deriving a grammar

A major drawback of the approaches analyzed so far was that symbols 'uttered' by agents were of a fixed length. This means that agents can only learn lexical mappings, but not grammars over an alphabet Σ, where each string is an element of Σ^*, i.e. strings will not *a priori* be determined to be of a certain length. However, since the occupation with grammar has been one of the major issues of twentieth century linguistics, I wondered whether the emergent evolution of syntax could be embedded in the framework of the present model.

One of the first problems then encountered is that any given neural network must have a fixed number of output units and that thus, any string produced by any of the agents is necessarily restrained to be not only finite but there must also be a fixed length $n \in \mathbb{N}$ which it *can* never exceed[7]. However, there *is* a way of modeling the evolution of grammar. If n was allowed to be large enough, then a grammar could be 'approximated' - analogously, for example, to a Taylor approximation in calculus - using the definition of $\Sigma^* := \Sigma^0 \cup \Sigma^1 \cup \ldots \cup \Sigma^n \ldots$, and cutting this infinite union off at some (large) n.

Having found a solution to the problem of approximating Σ^*, the next issue is to transform an output string $\in \Sigma^n$ [8] into one $\in \Sigma^k$, where $1 \leq k \leq n$, i.e., a method must be found for providing the ability for agents to utter variable-sized strings. A solution that suggests itself is the introduction of a 'special' symbol indicating the end of a sentence. Hence, the algorithm developed works as follows. When asked to verbalize a given meaning, an agent would generate a long (n characters long) sequence of characters, the preliminary string for this meaning. This string could include a special terminating symbol, referred to as *stop* symbol. All symbols after this stop were ignored in that the string was mapped to a string encompassing the symbols up to the stop and the stop itself, while the rest of the possible character positions was filled with stop symbols as well. The idea was that if an agent was to verbalize a, say, 3 bit meaning, it would be conditioned to produce a few 'valid' bits (possibly 3 as well) and then to produce at least one stop symbol. In doing

[7] Which presupposes the grammar to be regular.

[8] If - as it is done in this model - each bit in the net's output vector is interpreted as one character. Other authors have dealt with the issue differently (cf. Batali, 1998).

so, the agent could determine itself where it wanted the sentence to end, and so produce variable-length strings.

Of course, when introducing a third output symbol, the stop, the network's binary output function needs to be adapted. One could do this by either using a sigmoid output function and thus allowing continuous valued outputs that then have to discretized again, or, as it is done in my approach, allow a third output value by means of the following way. If an output unit is close enough to zero, say in the interval $[-\epsilon, +\epsilon]$, where ϵ is positive and small, it will be mapped onto the stop symbol and otherwise to $(+1)$ and (-1) in the usual way. Of course, it is necessary that agents cannot utter the empty string, λ, to denote a given meaning. Therefore, all but the first output unit must be subdued to this new output function.

This concludes the description of the basic algorithm which was complemented by a few modifications, some of which are to be reviewed in the following. It must be noted, beforehand, that diversity is not under investigation in the current analysis and therefore not implemented.

5.3.1 Basic Algorithm

Using this basic algorithm with some of the familiar parameters (e.g. agent number equal to 8, $t = 80$, 2^3 meanings, etc.), agent communication works fine and agents do evolve shared communication systems. However, having set the maximum number of output units (and thus the upper bound on string length) to 10, agents do not learn to produce short sentences. Repeating this experiment several times, it was found that agents used on average between 8 and 9 output units[9] although 3 to 4 would have sufficed to distinguish among the 8 different meanings. The initial hope that agents would learn to find the minimally necessary number of bits for adequate communication was thus not confirmed. Why is this so? Firstly, it must be stated that agents are not biased or restrained for minimality so that such an outcome could actually not be expected. Secondly, if agents mapped meanings to signals randomly then one would expect them to make use of 3 to

[9]That is, strings of length 8 or 9.

4 bits (that is, introduce the stop after three to four 'valid' characters)[10], so the rather large number of bits employed is in fact surprising. A reasonable explanation would be to assume that agents tend not to waste information *when they have the choice* because more information facilitates comprehension and - and this is important - the model makes the assumption that successful communication relies solely on the hearer and not on the speaker; it is the listener who is punished for not understanding the speaker, who, in turn, is indifferent to the listener's understanding him.

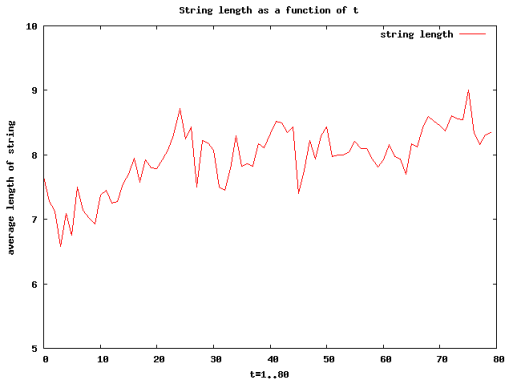

Figure 5.1: Basic algorithm. 8 agents, $m = 2^3$ meanings, $n = 2^{10}$ possible signals. String length as a function of $t = 1 \ldots 80$ (average over many generations). When agents are not biased for short strings, they will waste no information and strive instead for longer strings.

To conclude, since the basic algorithm does not force or incline agents to use 'short' signals, and strings therefore have (quasi) fixed length, the basic algorithm is not adequate for a simulation of the emergence of grammar, and has to be revised.

[10]If any of the two 'valid' characters plus the stop character were equally likely to be output for any given output unit, then the random variable Y that counts the position of the first stop symbol would have the expected value $\mathrm{E}[Y] = \sum_{k=2}^{n} \left(\frac{2}{3}\right)^{k-2} \cdot \frac{1}{3} \cdot k = 3.66$ for $n = 10$. However, while the two valid characters $(+1)$ and (-1) are indeed equally likely for any uniform initialization of a net's weights, the probability that an output bit takes on the value 'stop character' is very much dependent upon the size of ϵ; in fact, the smaller ϵ the more improbable it is that this character is output and the larger is the expected value of Y. In my experiments, with $\epsilon = 0.05$, Y was empirically found to be on average around 7.5 for net weights initialized uniformly from the interval $(-0.5, 0.5)$. Still, this is less than the value of 8 to 9 bits employed by one generation of agents *after* training has been completed, and so must be explained.

5.3.2 Modification 1

A first improvement, or rather biasing for 'short strings', was the idea to allow every agent a designated 'speaking time' that was chosen such that each agent could verbalize at least one meaning. The reasoning was that agents using short sentences were favored over those that used long sentences and that they could thus incite the others to do likewise. I was not particularly happy with this invention because it did not seem 'natural'. Was it the case that because some agents spoke 'concisely', their languages had thus better survival capabilities because in any given amount of time these agents could say 'more'? And thus influence any listener more profoundly? Interestingly, it appeared that such a modification had no effect upon the length of the strings uttered by the agents. On average, string lengths were about 8.1 bits per meaning, which was only a little less than the value of 8.5 bits achieved for the basic algorithm.
So this approach was abandoned, likewise.

5.3.3 Modification 2

Next, an enforcement of the employment of short strings was hoped to be accomplished by having speakers choose the shortest string $w \in \{a, b, STOP\}^+$ that they could interpret themselves as the given meaning. Formally, for a given meaning m a speaker would produce the sentence w that satisfies

$$\min_{|s|} f^{-1}(s) = m$$

where the norm $|\cdot|$ is defined to count the number of characters that are unequal to the stop symbol, starting from the first character; e.g., when $s = s_0 s_1 \dots s_{t-1} s_t \dots s_n$ with $s_i \in \{a, b\}$ for $i = 1, \dots, t-1$ and $s_t = STOP$, then $|s| = t$. In particular, the shortest string is then defined as the "shortest string counting from left to right". Hence, agents could only chop endings of words, not beginnings, or parts in the middle[11]. A

[11]This is solely in order to keep the model simple. I think, in natural languages all three kinds of abbreviations are possibly, as in *'Tis bitter cold* for *it is bitter cold*, German dialectal *g'stehn* for *gestehen*, and the droppings of the schwa phoneme, for example in Middle English words like *love, thilke*,

test of this modification yields the result that agents are now truly biased to utter short strings; the average string length is now 2.47 bits, with which agents can distinguish only $2^{2.47} = 5.5\ldots$ meanings. Because this is less than the minimally required three bits, it is foreseeable that agent communication is rather restricted. Figure 5.2 confirms this view; although, on average, agents improve their communicative abilities by learning, even after 80 epochs communicative success is not much larger than 50%, which is very poor. One might now conclude that the ability for agents to speak a grammar has cost the ability to acquire a *shared* grammar; it is possible however to increase t and thus restore mutual intelligibility in part. With $t = 180$, for example, communicative success was at an average final level of about 70%.

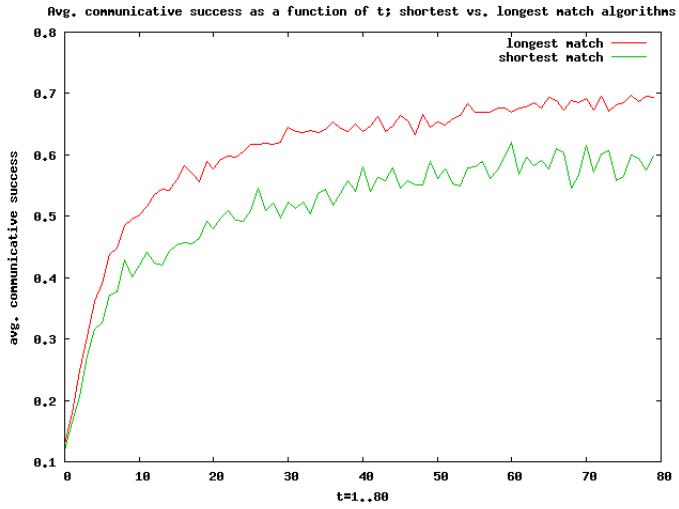

Figure 5.2: Avg. communicative success as a function of t. Depicted are the developments for the shortest left string algorithm (modification 2) and shortest right string algorithm (modification 3; top line). It must be noted that these lines are averages over many generations and that for some particular generation, it has also been observed that no shared language was established.

One of the languages derived by this process shall now be investigated, see Table 5.3.

Agents obviously have derived a VS language. The verb is encoded in the first position of the string uttered; the correlation is "0 = sleep $\leftrightarrow$ b" and "1 = eat $\leftrightarrow$ a". While this correspondence is completely regular, the rest of the correlations is fuzzier; however, the

etc.

Agent / meaning	1	2	3	4	5	6	7	8
000	baa	baa	bba	baa	bba	baa	bba	baa
001	aaa	a	ab	aa	a	a	aba	a
010	ba	ba	baa	ba	baa	ba	ba	ba
011	aa	aa	a	aaaba	aa	aa	aa	aa
100	bb	bb	bb	bbbab	b	bb	b	bb
101	a	ab	abb	a	abb	ab	a	ab
110	b	b	b	b	ba	b	bab	b
111	abb	abb	aab	abb	aab	abbb	aab	abb

Table 5.3: The grammars developed by the individual agents when agents produce strings according to the algorithm described under modification 2. The population consisted of 8 agents; 50 generations were chronicled. $\eta = 0.1$, $m = 2^3$, $n = 2^{10}$. The table presents the grammars established by the 50^{th} generation after $t = 180$ episodes of communication.

morphological analysis shown in Table 5.4 - as an average, so to speak, over the population - shall be suggested.

	I (=00)	*you* (01)	**she** (10)	we (11)
eat (=1)	a[a/b]a	a*a[a]*	a**b[b]**	a[b/a]b
sleep (=0)	b[a/b]a	b*a[a]*	b**b[b]**	b[b/a]b

Table 5.4: Morphological analysis of the grammar depicted in Table 5.3. Brackets denote facultative characters, a slash denotes 'free variation'. Actually, the table should be read in the following way: the character b in initial position has the meaning *sleep*, $a[a]$ in final positions has the meaning *you*, etc. Note the conventions for the interpretation of meanings set up in Table 5.1 (for example, *we-sleep* $\leftrightarrow$ 11-0). The analysis was found by inspection.

One final question has to be addressed here. Why is mutual intelligibility so much affected by the process of finding the shortest (left) string compatible with the understanding of the speaker? For one thing, it must be said that loss of information should naturally entail impairment of language performance. Additionally, the inclusion of a third output value into the neural network while retaining the given output function causes problems. This third value must - irrespective of its interpretation as a symbol character - take on the form of a numerical variable. It was arbitrarily set to 0.2 [12].

[12]The only value to be excluded was 0 because then no learning could have taken place, compare Chapter 3.

The problematic effect this has is that the algorithm updating the strength of connection between a net's input and output units cannot now differentiate between the numerical representation of the $STOP$ character, $(+0.2)$, and the numerical representation of the character b, $(+1)$. To see this, imagine that some listening agent incorrectly interpreted some position in the meaning vector it was supposed to decipher. To have a concrete example, assume an agent was supposed to understand the meaning $(+1)(+1)$ but instead understood $(+1)(-1)$. This agent will have to adapt the weights leading to the second input unit according to the formula given earlier and repeated here for the sake of convenience, $\Delta w_{ij} = \eta(x_i - x'_i)y_j$. The problem is now that the strength of the connection will be increased whenever y_j is $(+0.2)$ *or* $(+1)$. If an agent is repeatedly trained on the output $(+0.2)$, the strength of the connection will repeatedly be increased, making it less and less probable for the agent to output $(+0.2)$ [13] - because this only happens when the induced field is in the interval $[-\epsilon, +\epsilon]$ - but instead the agent will output $(+1)$.
There are still other factors involved. For example, it may always - also with only two output values - happen that an agent produces inconsistent output by mapping two distinct meanings to the same signal. Here, such a tendency is enforced when an agent g maps a certain meaning m to a certain signal s, but does not itself understand s to mean m but instead m'[14]. It will then occur that in order to utter m, g says s; on another occasion g may be prompted to utter m', which it will map, say, to s', where s' is longer than s but the two strings coincide on all positions of s (that is, s is a prefix of s'). Then it is possible that the shortest string having the meaning m' for agent g will also be s. So, this implementation of a third output value causes the introduction of additional noise.

5.3.4 Modification 3

The final and best-performing modification likewise searches the shortest string for a given meaning that an agent itself can still interpret as the meaning. However, it starts its search for this shortest string from the right end of the original string. Therefore,

[13] When asked to verbalize the meaning $(+1)$ at the given position.
[14] Although this is possible, such a phenomenon should be rather unusual, however.

the string found by this modification is always greater or equal to the string found when using the algorithm described above. To see this, assume an agent cannot understand the signals $s_o s_1 \dots s_k$, $k = 1, \dots, t-1$ as having the meaning m, but does understand the string $s_0 s_1 \dots s_{t-1} s_t$ to mean m. Then, whenever characters of the original string are chopped starting from the right - this time finding the first occurrence when the agent *does not* understand the shorter signal as meaning m - this search at the latest stops at position $t-1$. It can, of course, stop earlier, as when an agent might interpret, say, *aab* and *aabaa* as meaning m, but not *aaba*. In this case, the algorithm finding the shortest left string would choose the string *aab* (if all shorter strings were rejected), and the algorithm finding the shortest right string would choose *aabaa* as the 'correct' signal.

Under this implementation, agents used on average 3.36 bits, so that they could, in theory, distinguish between $2^{3.36} = 10.2 > 8$ different meanings. The grammars they derived are similar, if only longer, to the one discussed in the previous subsection.

5.4 Simulating sociolinguistic aspects of diversity

This work's last analysis explores the possibilities of integrating sociolinguistic factors in the computational formation of diversity, within the established framework. Among the model's parameters presented in Chapter 4, two appear to be promising candidates for being re-employed for the emulation of social processes; namely, the learning rate parameter η and σ, the neighborhood size[15]. In a hypothetical sociolinguistic environment with different social groups or classes, σ might represent the classes' 'wealth', their mobility, or their access to literature, equipment with televisions, mobile phones, etc. η might be regarded as the strength of bond among the members of a certain social class; how much do they identify with their group and thus with the language spoken by their fellow members? Can other dialects have influence upon them?

In a sample experiment, I devised a population hierarchy consisting of three classes[16],

[15] Although the role of σ as a possible sociolinguistic factor is stressed here, no experiments involving the parameter have been conducted in this work.

[16] This model is motivated by the class structures observed in occidental societies during the Middle

referred to as 'kings', 'citizens', and 'farmers' respectively. Farmers form the majority in number, amounting to 60% of the population, citizens comprise 30% of the total number of 120 agents, and kings 10%. Agents are placed randomly in their one-dimensional world, with the convention that 6 members of the same class always have to reside next to each other. This was done in order to reflect people's tendencies to join in groups (e.g. monks do not live separately within a society but instead gather to found a monastery), and in order to grant a chance to those groups with few members (e.g. kings) of spreading and maintaining their respective dialects. In order to keep the model simple, factors like 'convergence' to another social class, change in social population structure (for example, an increase or decrease of the number of the members of a certain class from one generation to the next), 'social' as well as 'geographical' mobility, were excluded. Also, society itself was not allowed to progress, i.e. for example by developing technology that allowed an increase (or decrease) of the neighborhood size over time.

In a first experiment, σ was set uniformly to 1.0 for all three classes. The respective values of η can be inferred from Table 5.5. It must be noted that there are now k^2 numbers of different η's, where k denotes the number of social classes in the modelled society; each class c_i has a η_{ij} for every other class c_j; this value represents class c_i's attitude towards class c_j. In particular, η_{ii} specifies the strength of bond among the members of class i.

class	king	citizen	farmer
king	0.25	0.05	0.025
citizen	0.2	0.15	0.05
farmer	0.2	0.15	0.1

Table 5.5: Learning rates of members of one class when listening to speakers of another class. This model was chosen so that the class of 'kings' was considered the most prestigious, the class of 'farmers' the least prestigious. The larger η_{ij}, the more willing will listeners of class i be to adapt to the speech of speakers of class j.

The results of the simulations are depicted in Figures 5.3 to 5.5. It can be observed

Ages or early modern times. One may not forget that other structures of social hierarchy might have been chosen just as well - such as, for example, a caste system analogous to those implemented in present day Hinduist societies where each social group is separated from the other groups (cf. Crystal, 1987: p.38).

that the simulation of sociolinguistic factors and the entailed inequality of individual speakers introduces further randomness into the model. For example, the number of dialects with only one speaker is very high in the setting under investigation (see Figure 5.4), indicating that some speakers are now influenced by different, possibly contradictory forces; geography competes with sociology. Speakers with such idiolects are found in all social groups. Finally, the resulting dialects mostly consist of speakers of the same class, which illustrates the newly-generated importance of 'class' for the emergence of variety; sometimes it is also found that farmers have adapted to the speech of their king(s) (or vice versa).

Altogether, however, the data assembled is not very conclusive and it is possible (or even likely) that enhancements have to be provided if the current model is to be able to account for sociolinguistic phenomena of linguistic diversity.

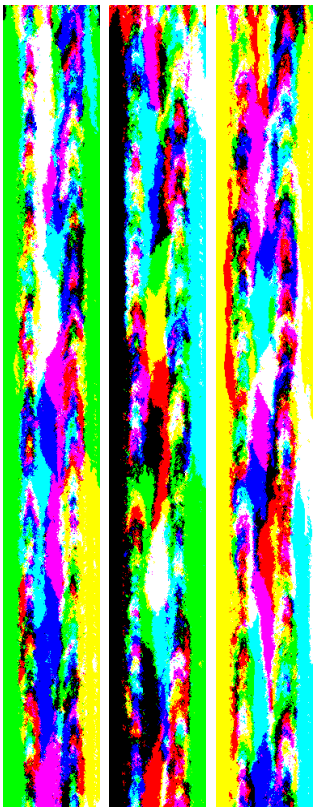

Figure 5.3: $t = 80$, 120 agents, 1000 generations, $m = 3$ meanings, $n = 2^3$ possible signals, $\sigma = 1.0$ for all members of all three classes, η_{ij} as in Table 5.5. For the distribution of social classes, see Figure 5.4.

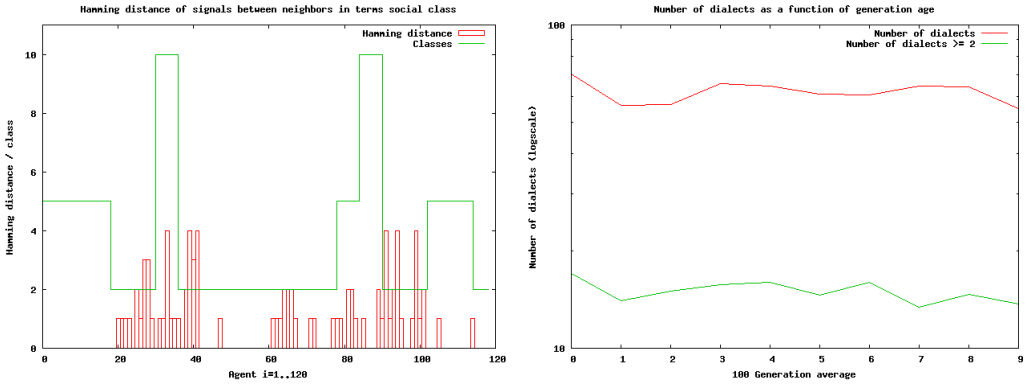

Figure 5.4: Analysis of the last generation of the simulation depicted in Figure 5.3. Left figure: hamming distance of signals used by neighboring speakers compared with change in social class. The top (green) lines represents social class: a value of 2 stands for the class of farmers, 5 for citizens, 10 for kings. It can be discerned that a change in class membership is accompanied by an increased tendency for change in signal use. Right figure: the top line depicts the evolution of the number of *all* dialects as a function of generation age, the bottom line displays the evolution of those dialects that have at least two members.

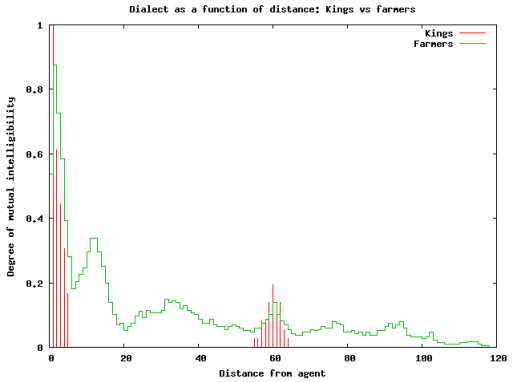

Figure 5.5: Analysis of the last generation of the simulation illustrated in Figure 5.3. Average degree of mutual intelligibility of speakers as a function of geographical distance, highest social class (kings, red line) vs. lowest social class (farmers, green line). One might argue that the finding that mutual intelligibility over geographical distance is slightly better for kings than for farmers is in accordance with the claim explicated by the social pyramid (see Chapter 2), namely that higher social status entails increased linguistic unity over spatial displacement. The fact that not all distances are covered for kings is derivative of their small number (12 kings vs 72 farmers).

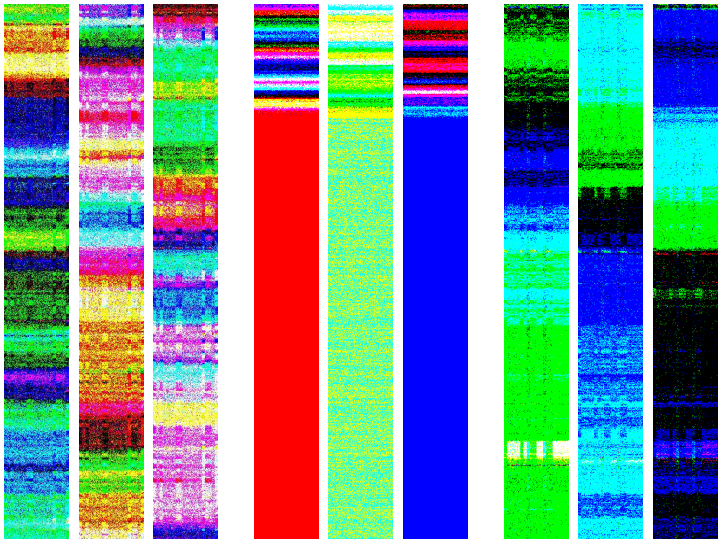

Figure 5.6: Can diversity be generated by sociolinguistic factors alone? Left figure: $t = 80$, 120 agents, social hierarchy as described in the text, $\sigma = \infty$ for all classes, $\eta_{ii} = 3.0$, $\eta_{ij} = 0.05$, for $i \neq j$. The figure gives no definite answer. Although change is present under these circumstances, and in each generation different dialectal groups have formed, variation seems more random and less systematic than in former experiments - remember, for example, that (at least 6) individuals of the same class are located in close vicinity to one another (see text) and that thus, a different pattern of diversity should be expected. Furthermore, that dialects contain members of all classes might indicate that diversity is owed to the large number of η_{ii}, and not to presence of different classes. However, two other experiments emphasize that the simulation of class structure has at least some bearing upon the emergence of change and variation: when classes were abandoned but η retained at the large value of 3.0 (for all *identical* agents), convergence to a global dialect is observed (figure in the middle). The figure on the right illustrates that a class hierarchy will preserve change even when η_{ii} is much smaller, namely $\eta_{ii} = 1.5$, again a situation in which agents have converged to a global dialect when no social structure was imposed upon their organization.

Chapter 6

Conclusions

> The important question is: What are the initial assumptions concerning the nature of language that the child brings to language learning, and how detailed and specific is the innate schema [...] (Chomsky, 1965: p.27)

This work was carried out in the tradition of the more recent approaches to the evolution of language arguing *against* a biological basis of the (syntactic) linguistic faculty (Batali, 1998; Kirby, 1998a; Kirby, 1998b; Hurford, 1998), and thus arguing against Chomsky (1965), Pinker (1994), and Pinker and Bloom (1990). It was, among others, the belief of these recent approaches that "biological natural selection is only one of the complex adaptive systems at work" (Kirby, 2001: p.2) in the process of language evolution, and that its contribution to this event was rather negligible. With biological evolution taking several time scales longer than, for example, *glossogenetic*[1] and ontogenetic evolution[2], a more productive approach to an explanation of the emergence of language might be found with the help of the latter two concepts. In any case, reason has to be supplied for why various grammatical principles should increase the probability of "fruitful sex" (Lightfood, 1991: p.69). Kirby (2001) consults, in contrast, glossogenetics as an *explanans* for the form of human grammar (see Figure 6.1).

Likewise, in the implementation conducted in this work, the availability of *complex* innate biases determining the shape of the grammars obtained was not required to generate, for example, compositionality[3]. It did not have to be explicitly programmed or

[1] By this, Kirby (2001) means the evolution of language itself.

[2] Language learning.

[3] However, the model still incorporated biases introduced by the learning rule, or, more general, the

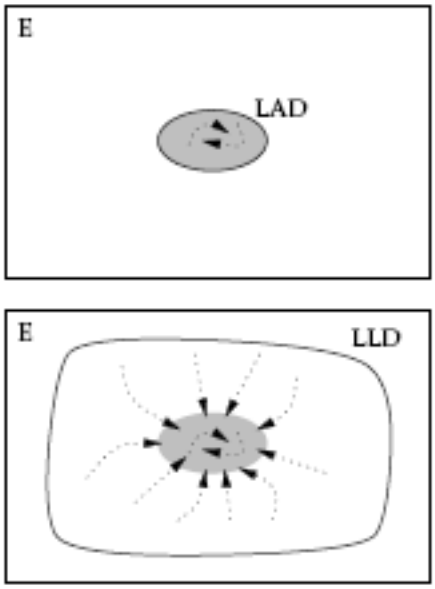

Figure 6.1: Two perspectives on the form of human grammars. E is the set of all logically possible languages. The top diagram illustrates the Chomskyan view that the innate language acquisition device (LAD) directly constrains the learner to learn the "humanly" possible languages. In the bottom diagram, the language learning device (LLD) is less constraining, and the particular characteristics of human languages are the result of a historical glossogentic evolution of languages in populations (after Kirby (2001): p.25)

hard-wired for the agents to 'generate' it in their course of language negotiation. Kirby (2001) claims that the same is valid for other prominent principles of the human linguistic capacity, such as recursiveness.

Still, not every aspect of language may be attributed to glossogenetic or ontogenetic

outline of the framework as a neural network. (Completely) bias free learning is futile (cf. Mitchell, 1995).

evolution, for otherwise it would not be unique to humans. In the current work, one of the prerequisites for agents to acquire a grammar (Chapter 5) was their ability to "use [their] own cognitive responses to predict those of others" (Batali, 1998: p.21); the signal an agent would send depended on what the agent itself would understand as the given meaning. Such an ability would not be classified as language-specific, but rather as general-purpose. Again, this fits in neatly with the claim made herein, namely, that Chomsky was wrong.

Before drawing too daunting conclusions, however, it must be noted that a general problem (referred to as 'scaling up' problem) of computational modeling of language evolution is that these approaches only deal with small subsets of the problems encountered in the real world. As Cangelosi and Parisi (2002) point out, "Most of the models simulate simple syntactic rules, such as subject-verb-object, not the rich morphological and grammatical features of human languages, and use lexicons with a few words as opposed to thousands of words comprising the lexicons of natural languages" (ibid.: p. 11). It might well be that processes that work fine for this subset may be inadequate to generate the same phenomena when it comes to more substantial issues.

One of the model's great advantages is its simplicity. With an almost trivial learning rule and a simple modification to this rule, it was possible to generate simple grammars. An efficient implementation thus allows this model to be a fruitful, virtual laboratory with which various questions of language evolution may be addressed. The idea to use an agent's own cognitive behavior to determine its sending performance was taken from Batali (1998), who, quite contrarily, employed a model that required (to my understanding) hours of computations before producing results comparable to the ones achieved in Chapter 5. To my knowledge, such a model has not yet been employed in the scientific community. As one of the model's unrealistic assumptions the separated 'evolution' of thought and language must be named. As Hutchins and Hazlehurst (2002) point out, "We find it implausible to look for the origins of language in interactions where fully composed meanings are injected into the mind of a listener before a public expression of that meaning is encountered". What is required, instead, is a scenario where language

and thought can coevolve, for example, as is done in Steels (2002).

Other aspects where future work might add on is the employment of different topologies wherein agents interact. This is by no means a trivial question because, for instance, a symmetric random walk, another exemplar of a highly stochastical process, is recurrent in one and two dimensions, but not so in three (and further) dimensions (Georgii, 2004: p.169f.). It might also be possible to endow agents with life cycles longer than two periods and, for example, adapt their learning rate parameter and the neighborhood size according to the cycle an agent currently finds itself in. Infants, for example, will in their early years even in highly-industrialized countries with a high degree of mobility primarily have their parents as communication partners. Given the stochasticity of the processes involved this is, again, possibly not just a trivial expansion of phenomena investigated earlier. Finally, appending to the current work a thorough mathematical analysis with respect to dynamical systems may supply further valuable insight. It is not unlikely that dynamical systems are the only way of accounting for the miracle of language origin and evolution that has puzzled linguists and scholars since Biblical times.

Acknowledgments

I want to thank Christopher Campbell, Anton Davydov, and Hendrik Niederlich for the ungraceful but indispensable labor of proofreading this work. I thank Leo (http://dict.leo.org) for enriching my English vocabulary.

Bibliography

[1] Batali, J. (1998). Computational simulations of the emergence of grammar. In: Hurford, J. R., Studdert-Kennedy, M. and Knight C., (eds.), *Approaches to the Evolution of Language: Social and Cognitive Bases*, Cambridge: Cambridge University Press.

[2] de Boer, Bart (2002). Evolving Sound Systems. In: Angelo Cangelosi and Domenico Parisi (eds.), *Simulating the evolution of language*, pp. 79-97. London: Springer.

[3] Cangelosi, A., (1999). Modeling the evolution of communication: From stimulus associations to grounded symbolic associations. In: D. Floreano, J. Nicoud, and F. Mondada, (eds.), *Advances in Artificial Life (Proceedings ECAL99 European Conference on Artificial Life)*, Berlin: Springer-Verlag, pp. 654-663.

[4] Cangelosi, A., and Parisi, D. (2002). Computer simulation: A new scientific approach to the study of language evolution. In: Cangelosi, Angelo and Parisi, Domenico, (eds.), *Simulating the Evolution of Language*, pp. 3-28. London: Springer Verlag.

[5] Chambers, J.K. and Trudgill, P. (1980). *Dialectology*. Cambridge: Cambridge Univ. Press.

[6] Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

[7] Christiansen, M. H., Dale, R., Ellefson, M. R., and Conway, C. M. (2002). The role of sequential learning in language evolution: Computational and experimental studies. In: Angelo Cangelosi and Domenico Parisi, (eds.), *Simulating the Evolution of Language*, pp. 165-188. London: Springer Verlag.

[8] Crystal, D. (1987). *The Cambridge encyclopedia of language*. Cambridge: Cambridge Univ. Press.

[9] Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und Philosophische Kritik*, 100: pp. 25-50.

[10] Georgii, H.-O. (2004). *Stochastik*. Berlin: de Gruyter.

[11] Haykin, S. (1994). *Neural networks - A comprehensive foundation*. Macmillan.

[12] Hurford, J. (1998). Social transmission favours linguistic generalisation. In: C. Knight, J. Hurford, and M. Studdert-Kennedy (eds.), *The Emergence of Language*.

[13] Hutchins E., Hazlehurst B. (1995). How to invent a lexicon: the development of shared symbols in interaction. In: Gilbert N, Conte R (eds.), *Artificial Societies: The computer simulation of social life*. UCL Press.

[14] Hutchins, E. and Hazlehurst, B. (2002). Auto-Organization and Emergence of Shared Language Structure. In: Angelo Cangelosi and Domenico Parisi, (eds.), *Simulating the Evolution of Language*, pp. 279–306. London: Springer Verlag.

[15] Kirby, S. (1998a). Language evolution without natural selection: From vocabulary to syntax in a population of learners. In: C. Knight, J. Hurford, and M. Studdert-Kennedy (eds.), *The emergence of language*.

[16] Kirby, S. (1998b). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In: C. Knight, J. Hurford, and M. Studdert-Kennedy (eds.), *The Emergence of Language*.

[17] Kirby, S. (2001) Learning, bottlenecks and the evolution of recursive syntax. In: Briscoe, EJ (ed.), *The evolutionary emergence of language: Social function and the origins of linguistic form*, pp. 303-323. Cambridge: Cambridge University Press.

[18] Kirby, S., and Hurford, J. (2002). The emergence of linguistic structure: an overview of the iterated learning model. In: Cangelosi, Angelo; Parisi, Domenico (eds.), *Simulating the evolution of language*, pp. 121-148. London: Springer Verlag.

[19] Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

[20] Lass, Roger (1997). *Historical Linguistics and Language Change*. Cambridge: Cambridge Univ. Press.

[21] Lightfood, D. (1991). Subjacency and sex. *Language and communication*, 11(1): 67-69.

[22] Livingstone, D. (2002). The Evolution of dialect diversity. In: Angelo Cangelosi and Domenico Parisi, (eds.), *Simulating the evolution of language*, pp. 99-118. London: Springer Verlag.

[23] Livingstone, D. and Fyfe, C. (1999). Modeling the Evolution of Linguistic Diversity. In D. Floreano, J. Nicoud and F. Mondada, (eds.), *ECAL99*, pp. 704–708. Berlin: Springer-Verlag.

[24] Noble, J., Di Paolo, E. A. and Bullock, S. (2002). Adaptive factors in the evolution of signalling systems, In: Cangelosi, A. and Parisi, D., (eds.), *Simulating the Evolution of Language*, pp. 53-78. London: Springer.

[25] Pinker, S. (1994). *The language instinct*. Penguin.

[26] Pinker S., Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13: pp. 707-784.

[27] Steels, L. (2002). Grounding Symbols through Evolutionary Language Games. In: Angelo Cangelosi and Domenico Parisi, (eds.), *Simulating the Evolution of Language*, pp. 211-226. London: Springer Verlag.

[28] Steels, L., Kaplan, F., (1998). Stochasticity as a source of innovation in language games. In: Adami C, Belew RK, Kitano H, Taylor CE (eds.), *Proceedings of Artificial Life IV*, MIT Press.

[29] Wardhaugh, R. (2002). *An Introduction to Sociolinguistics*. Malden, Oxford, Melbourne and Berlin: Blackwell.

[30] UCL Dept. of Phonetics & Linguistics, October 10, 2006. Retrieved from http://www.phon.ucl.ac.uk/home/johnm/sid/isogloss.htm.

[31] Wikipedia - The free encyclopedia, October 10, 2006. Retrieved from http://en.wikipedia.org/wiki/Rhotic_and_non-rhotic_accents.

[32] Centre for English Language Studies, October 10, 2006. Retrieved from http://www.cels.bham.ac.uk/resources/essays/macedo6.pdf.