To be presented at:
Language Evolution and Computation Workshop/Course at the
European Summer School in Logic Language and Information,
Vienna, 25-29 Aug 2003

Creole Viewed from Population Dynamics

Makoto Nakamura*[†], Takashi Hashimoto[‡] and Satoshi Tojo[†]
Graduate School of {[†]Information, [‡]Knowledge} Science,
Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Tatsunokuchi-machi, Nomi-gun, Ishikawa, 923-1292, Japan

Abstract

Creole is one of the main topics in various fields concerning the language origin and the language change, such as sociolinguistics, the developmental psychology of language, paleoanthropology and so on. Our purpose in this paper is to develop an evolutionary theory of language to study the emergence of creole. We discuss how the emergence of creole is dealt with in the perspective of population dynamics. The proposal of evolutionary equations is a modification of the language dynamics equations by Komarova et al. We show experimental results, in which we could observe the emergence of creole. Furthermore, we analyze the condition of creolization in terms of similarity among languages. We conclude that a creole becomes dominant when pre-existing languages are not similar to each other and rather similar to the newly appeared language (would-be-creole); however the new language must not be too similar, in which case pre-existing languages remain and coexist.

Keywords: Population Dynamics, Creole, Similarity among Languages, Language Dynamics Equations

1 Introduction

Generally, all human beings can learn any human language in the first language acquisition. One of the main purposes of language use is to communicate with others. Therefore, it is easy to consider that the language learners come to obtain the language which they hear most in the community, i.e., in most cases, children will develop their parental languages correctly. When people do not have a common language to communicate with each other, such as plantation economies, slave trade, and so on, they come to use a simplified language called *pidgin* to bridge communication gaps between speakers of mutually unintelligible languages. After that, the children of the pidgin speakers may obtain a full-fledged new language called *creole* as their native language [6]. Thus, children have an ability to learn and create the most communicative language in the community.

In the stream of simulation studies of language evolution [4], the emergence of creole is also studied [10]. Briscoe [3] has reported sophisticated models of human language acquisition by means of a multi-agent model. However, because the number of agents was finite, the results were often hard to be generalized to explain general phenomena in the real world, from which the most multi-agent models had suffered.

To overcome this drawback of multi-agent models from a different viewpoint, Nowak et al. developed mathematical theory of the evolutionary dynamics of language [13]. By defining similarity and payoff between languages, based on the assumption of the universal grammar, Komarova et al. [8]

proposed *language dynamics equations* in which the transition of population among finite number of languages described by differential equations. However, in the framework of evolutionary dynamics of language, the emergence of creole was not discussed yet.

Our purpose in this paper is to develop the evolutionary theory of language in order to investigate the emergence of creole. We have already seen creolization by introducing the assumption that language acquisition of children is affected both by the distribution of population and by the exposure rate to other languages than their parental one [9]. In this paper, we analyze the condition of relationship among languages for creole to emerge and to be dominant.

In Section 2, we discuss how we consider creolization in the context of population dynamics. In Section 3, we describe the language dynamics equations and our modification of the equations. Section 4 reports our experiments. We present a discussion and a conclusion in the last two sections.

2 Creolization in Population Dynamics of Language

In this section, we describe creole from the viewpoint of population dynamics. We showed the emergence of creole in population dynamics of language [9], which is caused by transition of population among grammars and the exposure probability of children. Here, we discuss how the emergence of creole is considered in population dynamics.

2.1 Creole and Population Dynamics

We presuppose that the emergence of creole strictly depends on the population distribution, as opposed to traditional linguistic explanations [2, 6]. From the viewpoint, a creole is considered as such a grammar G_c that; A) $x_c(0) = 0$, $x_c(t) > \theta_c$ or B) $x_c(0) = 0$, $x_c(t) > \theta_d$, where $x_c(t)$ denotes the distribution of the population of G_c at time t, and θ_c and θ_d denote certain thresholds to be regarded as *coexistent* and *dominant*, respectively. These definitions represent that some individuals come to speak a language that no one spoke at the initial state, and consequently, A) a fixed number of individuals keeps the grammar, and B) the distribution of the language speaker occupies the most in the community.

2.2 Similarity among Languages

The S matrix in population dynamics denotes the similarity between grammars, which is determined by the probability s_{ij} that a speaker who uses a grammar G_i will say a sentence that is understandable by speakers of another language G_j; thus, each of $S = \{s_{ij}\}$ is a constant diachronically. Generally, the S matrix is uniquely calculated when the grammars and the probability for each sentence are given. Suppose that an individual who uses G_i utters a sentence in $L(G_i)$ with the uniformly same probability, then s_{ij} is the number of common sentences between $L(G_i)$ and $L(G_j)$ divided by the number of sentences in $L(G_i)$. Therefore, diagonal elements of the S matrix are always 1. Under the assumption, Fig. 1 shows the relationship among languages in the S matrix. The shaded part in the figure denotes that $L(G_1)$ and $L(G_2)$ share common sentences. In this case, s_{12} is greater than s_{21}, because the common part is rather small in $L(G_1)$ than in $L(G_2)$. The size of $L(G)$ concerns the generative capability of the grammar. Because the power of expressiveness is considered to be similar among languages, we should regard that the size of $L(G_i)$'s are same and that s_{ij} is nearly equal to s_{ji}. Thus, the S matrix should be an approximately symmetrical matrix.

In the above discussion, it is assumed that the member of conceivable grammars is finite and predefined. In this sense, creole is also included in them and has the similarity with the other languages

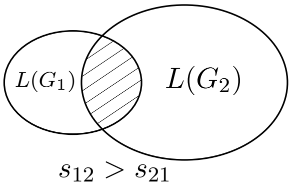


Fig. 1 The relationship among languages in the S matrix

in the S matrix. This is justified by the perspective of the universal grammar. We presuppose that creole may occur according to the similarity to the other languages, and thus we study the conditions of the similarity for creolization.

3 Population Dynamics of Grammar Acquisition

We introduce modified language dynamics equations after reviewing Komarova et al. [8]'s original ones.

3.1 Komarova et al.'s Language Dynamic Equations

Komarova et al. [8,13] proposed a mathematical theory for the evolutionary and population dynamics of grammar acquisition. In their model, given the principles in the universal grammar, the search space for candidate grammars is assumed to be finite, that is $\{G_1, \ldots, G_n\}$. Let $x_j(t)$ be the ratio of the population of G_j speakers, where $\sum_{j=1}^{n} x_j(t) = 1$. Thus, the model is defined in population dynamics in which individuals change their own grammar from generation to generation. The language dynamics equations are mainly composed by (i) the similarity between languages as the matrix $S = \{s_{ij}\}$ and (ii) the probability that children fail to acquire their parental language as the matrix $Q = \{q_{ij}\}$. Individuals reproduce children, the number of which is determined by the *fitness* such as: $f_i(t) = \sum_{j=1}^{n}(s_{ij} + s_{ji})x_j(t)/2$. The language dynamics equations are given by the following differential equations:

$$\frac{dx_j(t)}{dt} = \sum_{i=1}^{n} q_{ij} f_i(t) x_i(t) - \phi(t) x_j(t) \qquad (j = 1, \ldots, n), \tag{1}$$

where $\phi(t) = \sum_{i=1}^{n} f_i(t) x_i(t)$ and the term '$-\phi(t)x_j(t)$' makes the total population size keep constant.

In those equations, the fitness f_i for each grammar is regarded as its communicability, which represents a probability that a sentence uttered by an individual is recognized in the community. Total distribution of children of G_i speakers becomes $f_i x_i$. By the definition of the Q matrix, children are allowed to make mistakes during language acquisition. It is possible for a child to learn grammar G_i from her parents and to end up speaking grammar G_j. The probability of such transition is defined as $Q = \{q_{ij}\}$. In their work, it is also assumed that only adult individuals talk to the other language groups, while children communicate with only their parents. In this circumstance, it may be difficult to consider that the children mistake their parental grammar for another one.

3.2 Niyogi's Model

Niyogi [11, 12] gives actual examples of the Q matrix with linguistically well-grounded grammars together with the trigger learning algorithm (TLA) [7]. However, there is an unrealistic Markov structure which implies that some children cannot learn certain kinds of language, as we pointed out in [9].

3.3 Our Modification

Thus far, we have modified the language dynamics equations to include some constraints concerning transition among languages [9]. We have shown by computer simulations [10] that the population could be changed when children are exposed not only to their parental language but also to other languages. It is reasonably supposed that the transitions depend on the distribution of population of languages for children to be exposed. Therefore, the Q matrix should change through generations. Our prime revision is to introduce the probability α that children are affected by the other language speakers than their parents. We call α the *exposure probability*. A child hears not only parental language but also other languages in proportion both to the rate of the exposure α and to the distribution of population of grammars (See Fig. 2(a)). The probability which the children learn a language from their parents comes to $(1-\alpha)$. Note that α does not exclude children's parental language; it is also included in α in proportion to the distribution of population as well as the other languages.

Since the distribution of population changes in time, the Q matrix should include the time parameter t, that is, Q is redefined as $\overline{Q}(X(t)) = \{\overline{q}_{ij}(t)\}$, where $X(t) = (x_1(t), x_2(t), \ldots, x_n(t))$. We call $\overline{Q}(X(t))$ the *modified accuracy matrix*. Together with the S matrix and a given α, a learning algorithm determines $\overline{Q}(X(t))$. Thus, the new language dynamics equation is as follows:

$$\frac{dx_j(t)}{dt} = \sum_{i=1}^{n} \overline{q}_{ij}(t) f_i(t) x_i(t) - \phi(t) x_j(t) \qquad (j = 1, \ldots, n). \tag{2}$$

3.4 The Learning Algorithm

We introduce a simple learning algorithm which resolves Niyogi [11]'s problem mentioned above. The learning algorithm becomes as follows (See also Fig. 2(b)):

1) In a child's memory, there supposed to be a score table of grammars.

2) The child receives a sentence uttered by an adult.

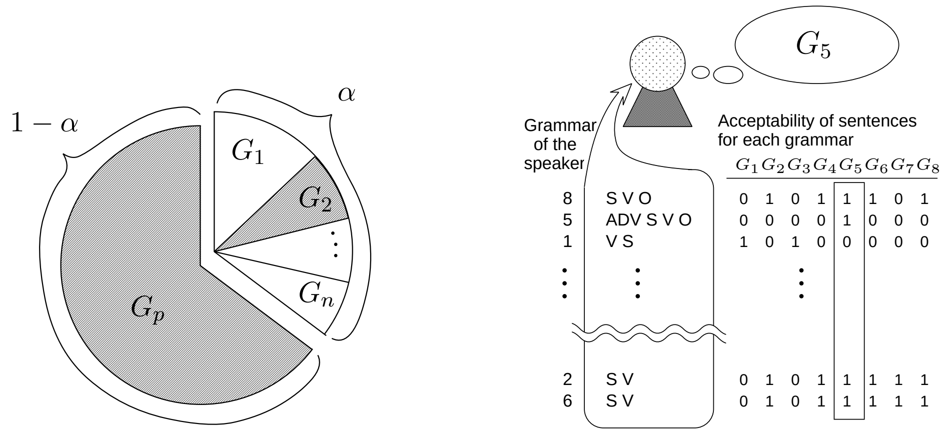
3) For each grammar, if a sentence is acceptable for the child, the grammar scores a point in her memory.

4) 2) and 3) are repeated until the child receives a fixed number of sentences that is regarded as enough for the estimation of the grammar.

5) The child adopts the grammar with the highest score.

Here, we introduced the exposure probability α that prescribes the ratio a child talks to people other than her parents. Thus, the estimated grammar of the child is G_{j^*} such that:

$$j^* \;=\; \underset{j}{\operatorname{argmax}} \{\alpha \sum_{k} s_{kj} x_k(t) + (1-\alpha) s_{pj}\}. \tag{3}$$



(a) The exposure probability α ($p = 2$)

(b) The simplest learning algorithm

Fig. 2 Introducing the exposure probability and the learning algorithm

From the learning algorithm, we give the modified accuracy matrix $Q(X(t)) = \{\overline{q}_{ij}(t)\}$ in [9] as follows:

$$\overline{q}_{ij}(X(t)) = \frac{(\alpha \sum_k s_{kj} x_k(t) + (1-\alpha) s_{ij})^{n-1}}{\sum_l (\alpha \sum_k s_{kl} x_k(t) + (1-\alpha) s_{il})^{n-1}}. \quad (4)$$

4 Experiments

In this section, we show the experimental result of the language dynamics equations of population-based transition in Section 3. We examine the conditions that creole appears and comes to be dominant in combinations of the S matrix.

4.1 Settings

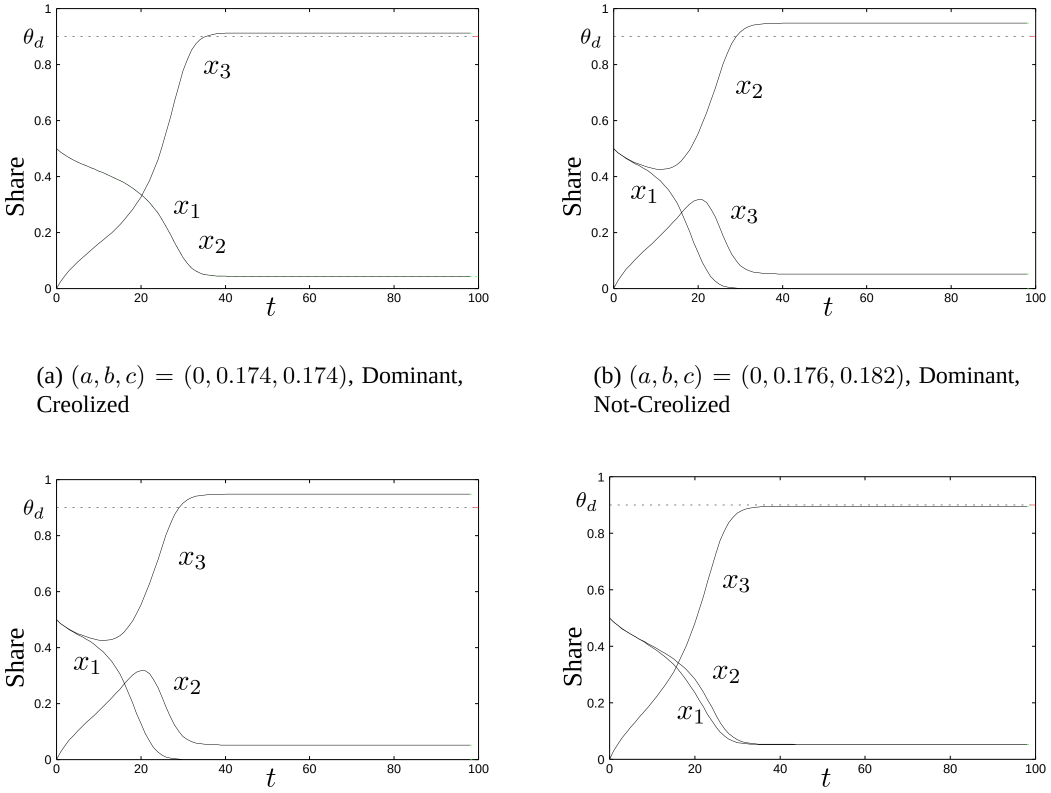
Here, we give parameters for the experiments. Since it is clear that creolization is the most observable in case $\alpha = 1$, we examine this case through the experiments. We analyze the case of three languages with the symmetry in s_{ij} and s_{ji} from the reason explained in Section 2.2, that is, the S matrix is formed as below:

$$S = \begin{pmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{pmatrix}. \quad (5)$$

The initial populations are given as $x_1(0) = x_2(0) = 0.5, x_3(0) = 0$. Therefore, we parametrize a, b and c in Eqn (5), and then research the mutual dependency in which G_3 becomes creole.

4.2 Conditions of Creole to be Dominant

The experiment aims at finding boundaries in the parameter space as to which language would be dominant. We refer to this situation as *dominant creolization*. Fig. 3(a) shows that the creole G_3 is dominant, in which the threshold for a language to be dominant is defined as $\theta_d = 0.9$. The S matrix is



(a) $(a, b, c) = (0, 0.174, 0.174)$, Dominant, Creolized

(b) $(a, b, c) = (0, 0.176, 0.182)$, Dominant, Not-Creolized

(c) $(a, b, c) = (0, 0.182, 0.176)$, Dominant, Not-Creolized

(d) $(a, b, c) = (0, 0.188, 0.189)$, Not-Dominant, Creolized

Fig. 3 The relationship between dominant creole and the S matrix

set to $(a, b, c) = (0, 0.174, 0.174)$, that is $b = c$. The value of $a = 0$ denotes that there is no common sentence in G_1 and G_2. Because the languages G_1 and G_2 play same roles, the dynamics of x_1 and x_2 are completely the same. In the figure, the language G_3 which no one spoke at the initial state comes to occupy the population with the rate of more than θ_d, while x_1 and x_2 declined concurrently. Namely, this is the emergence of a dominant creole in population dynamics.

When the values b and c increases slightly, the dominant language changes to another one while the share of creole G_3 is getting smaller. Fig. 3(b) represents the dynamics with the S matrix set to $(a, b, c) = (0, 0.176, 0.182)$. The figure denotes that G_2 becomes dominant, while G_1 eventually disappeared though it had the same population with G_2 at the initial state. When we transposed the value of b and c as $(a, b, c) = (0, 0.182, 0.176)$, the dynamics does not change but the dominant language is replaced (See Fig. 3(c)).

Changing the values of b and c continuously, we observed the sheer boundary of the change of the dominant language between them. Fig. 4 shows that the boundaries for the creole (G_3) to be dominant for several values of a. The crosses (×) in the figure represent the parameter values corresponding to Fig. 3(a)–(d), respectively. Because the parameters b and c work similarly G_1 and G_2, the boundaries are symmetric along the line $b = c$. In the figure, the long curve of the outmost boundary ($a = 0.00$) intersecting between (a) and (b) in Fig. 4 stands for the boundary of the change of the dominant

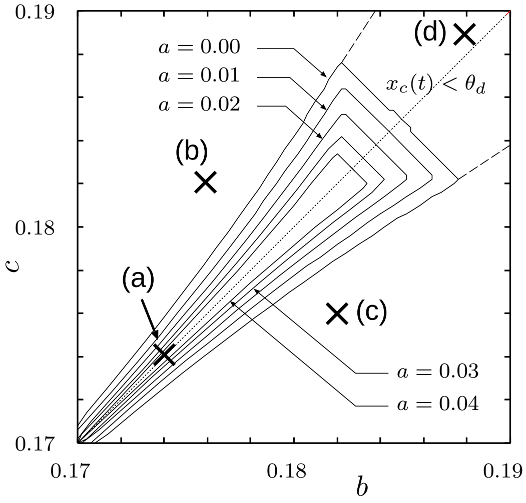


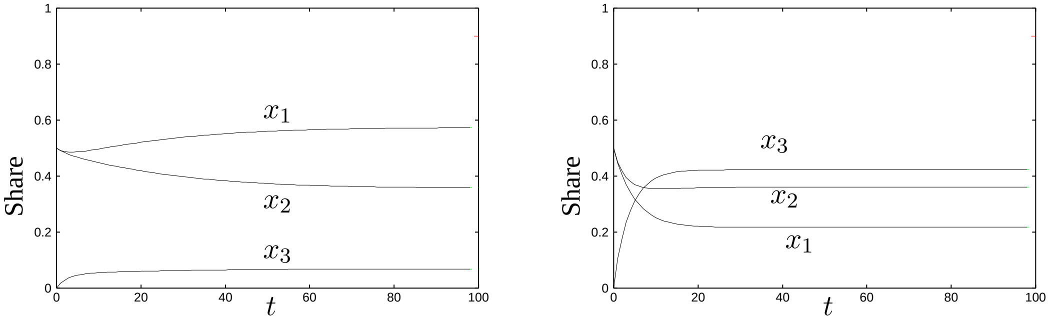
Fig. 4 Conditions for Dominant Creole ($\theta_d = 0.9$)

language. Inside of the lines, G_3 is the dominant language (Fig. 3(a)), above of the upper line, G_2 is dominant (Fig. 3(b)) and below of the lower line, G_1 is (Fig. 3(c)). Even if the threshold for dominant language, θ_d, were eased to be lower, this boundary had not changed. It is also the case with the different value of a. Thus, the long side of boundaries among dominant language is independent of θ_d for a given value of a. The broken lines in Fig. 4 are the boundaries of the dominant creolization for smaller values of θ_d at $a = 0.00$.

Next, we consider the short side of the boundaries in Fig. 4, that is the line crossing perpendicularly to the line $b = c$ (dotted line). This also represents critical conditions whether creole occurred or not for several values of a. These boundaries are, however, different from the one mentioned above. In Fig. 3(d) with $(a, b, c) = (0, 0.188, 0.189)$, we observed that G_3 still remained as the most populous language although the rate x_3 was a little less than $\theta_d = 0.9$. If θ_d was eased to lower, G_3 at the parameters of (c) would be regarded as creole. Hence, the position of the short line can shift along the line $b = c$ with the value of θ_d. It is easy for us to recognize that higher θ_d shrinks the area of creolization in the parameter space and vice versa.

As the larger the values b and c, the larger population transfer from G_1 and G_2 to G_3, respectively, and the width between the upper and lower boundaries grows. At the same time, however, $x_3(t)$ converges to smaller values at $t \to \infty$ with larger b and c. At last, $x_3(t \to \infty)$ falls short of θ_d at the short side boundaries in Fig. 4. This is because the more population shifts from G_3 to G_1 and G_2 by the larger values of b and c. Inversely, for the smaller b and c, say $b \approx c \lesssim 0.135$, in spite of the large share of x_3, the time needed for G_3 to dominate the all population comes to be longer that we could not observe further creolization.

To observe further details of the region of creole, we parametrized a in Eqn (5). In Fig. 4, regions of creole come to narrow with increasing a. Since a large value of a promotes communicability between G_1 and G_2 and enlarges the transition between them, no large population shifts from them to G_3. Therefore, the increase of a results in no dominant creolization.



(a) $(a, b, c) = (0.35, 0.25, 0.1)$, Coexistent, Not-Creole

(b) $(a, b, c) = (0.3, 0.4, 0.5)$, Coexistent, Creole

Fig. 5 Coexistent-language set

4.3 Summary of the Results

The conditions of the off-diagonal elements a, b and c in the symmetric similarity matrix S for the emergence of dominant creole is:

$$a \lesssim 0.1 \tag{6}$$

$$0.13 \lesssim b \simeq c \lesssim 0.2 \tag{7}$$

In this range, changes of a, b and c result in the followings:

1) When b and c are large, a must be small; in which case b and c might much differ. In this case, the share rate of G_3 becomes rather small at the time of convergence.

2) On the contrary, when a is small enough, b and c should be small. In this case G_3 dominates and converges in a short period.

5 Discussion

5.1 Conditions of Creolization in Natural Language

We obtained the condition of similarities among languages, in which a creole emerges and is to be dominant. Let us consider what this condition implies in the context of natural language. Suppose two languages, say super-stratum and sub-stratum languages. The condition $a \lesssim 0.1$ (Eqn (6)) indicates that these two languages are not similar. The two languages must be less similar to each other than to creole. If they are similar enough, the communication gap between users of these two languages is not so wide that they can understand each other to some extent. Thus, no pidgin or creole is needed. The condition, Eqn (7), says that the values b and c should not be too small but should be relatively small. If the pre-existing languages are similar enough to the creole, that is, the second in equality of the condition, Eqn (7), does not hold, the creole emerges but the users of the pre-existing languages can communicate with the creole users, then the speakers of the pre-existing languages do not diminish. The former part of the condition, Eqn (7), means that when a newly appeared language has no similarity to two pre-existing languages, it hardly becomes a creole.[1] Eqn (7) also confines the similarity of the pre-existing languages to the creole within a narrow range ($b \simeq c$). When the similarity of creole to the super-stratum language is enough larger than that of to the sub-stratum language, the super-stratum language comes to be dominant, and vice versa.

5.2 Language and Dialect

In this paper, we thoroughly analyzed the parameter region at which creole is dominant. When we look at the whole parameter space, we found the following four categories about dominance and creolization:

i) Dominant and Creolized; like Fig. 3(a)

ii) Dominant and Not Creolized; like Fig. 3(b)

iii) Coexistent (no dominant language) and Creolized like; Fig. 5(a)

iv) Coexistent and Not Creolized; like Fig. 5(b)

According to our preliminary investigation, the parameter region of the coexistent categories (iii) and iv)) is $a, b, c \gtrsim 0.3$, where the similarities among languages are relatively high and at this rate the language users can communicate with each other to some extent. This situation is better to be regarded as dialects rather than different independent languages. There is, in general, no clear boundary between dialects in a language and different languages from the pure-linguistic viewpoint.[2] From our results, the similarity of 0.3 may be a rough criterion for dividing between them.

6 Conclusion

In this paper, we argued that the emergence of creole in population dynamics of languages and showed that the emergence is affected by the similarity among languages as well as the distribution of population of the languages in the community. We obtained results for the condition of the similarity for dominant creolization as follows.

A) The pre-existent languages are not similar to each other, but to the newly appeared language.

B) The newly appeared language must not be too similar to the pre-existent languages. Otherwise, the pre-existent languages remain and coexist.

C) The pre-existent languages have approximately same distance to the newly appeared language with regard to similarity.

Creolization has not been dealt from the viewpoint of population dynamics and similarity among pre-existing and a creole, although similarity among creoles has been investigated [1]. Our contribution is to address a prediction about similarity among languages for creole to develop. This prediction should be tested empirically by observing grammars of various creoles and their original super- and sub-stratum languages.

[1] Since the similarity here is not the extent of the mixture of grammars in two languages, this implication does not contradict the fact that grammar of a creole is not a blend of those of super- and sub-stratum languages.

[2] The boundary is often settled politically such as Serbian, Croatian and Bosnian in Bosnia and Herzegovina. [5]

We argued the relationship between a language and dialects. Since the difference between language and dialect concerns the grammatical features, it is not possible to distinguish them only with the similarity, much less creole. This is an important problem in the present population dynamics. Therefore, further progress is needed to develop linguistic features into the population dynamics. We need to study in further generalized and actual conditions, to clarify the boundary conditions of creolization.

References

[1] Bickerton, D.: Roots of Language, Karoma Publishers, Ann Arbor, MI (1981)

[2] Bickerton, D.: Language and Species, The University of Chicago Press, Chicago (1990)

[3] Briscoe, E.J.: Grammatical Acquisition and Linguistic Selection, In: Briscoe, T. (ed.): Linguistic Evolution through Language Acquisition, Cambridge University Press, Cambridge (2002) pp.255-300

[4] Cangelosi, A., Parisi, D. (eds.): Simulating the Evolution of Language. Springer, London (2002)

[5] Comrie, B., Matthews, S., Polinsky, M.: The Atlas of Languages Quatro Publishing, London (1996)

[6] DeGraff, M. (ed.): Language Creation and Language Change, The MIT Press, Cambridge, MA (1999)

[7] Gibson, E., Wexler, K.: Triggers, Linguistic Inquiry, **25** (1994) pp.407-454

[8] Komarova, N.L., Niyogi, P., Nowak, M.A.: The Evolutionary Dynamics of Grammar Acquisition, J.Theor.Biol. **209** (2001) pp.43-59

[9] Nakamura, M., Hashimoto, T., Tojo, S.: The Language Dynamics Equations of Population-based Transition – a Scenario for Creolization –, Proceedings of the 2003 International Conference on Artificial Intelligence (IC-AI'03), CSREA Press (2003) (to appear)

[10] Nakamura, M., Tojo, S.: The Emergence of Artificial Creole by the EM Algorithm, Proceedings of the Fifth International Conference on Discovery Science (DS2002), Lecture Notes in Computer Science 2534, Springer (2002) pp.374-381

[11] Niyogi, P.: The Informational Complexity of Learning from Examples, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA (1994)

[12] Niyogi, P., Berwick, R.: The logical problem of language change. Technical Report AI Memo 1516 / CBCL Paper 115, MIT AI Laboratory and Center for Biological and Computational Learning, Department of Brain and Cognitive Sciences (1995)

[13] Nowak, M.A., Komarova, N.L.: Towards an evolutionary theory of language, Trends in Cognitive Sciences **5**(7) (2001) pp.288-295