R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 176-192.*

Logistic regression model for predicting language change

Shoichi Yokoyama
Haruko Sanada

1. Introduction

Scientific studies, including language-related research, are concerned with phenomena of changes and the mechanisms behind the phenomena. Research in other natural science has shown that phenomena, in which choices and changes are involved, can be efficiently represented by s-shape curves and by corresponding logistic regression models in biological and social research, such as studies on the increase of animals in population ecology and increasing popularity of new products in economics. In fact, research has said that language changes are often represented by S-shaped curves, which starts with little and slow changes at the beginning, fast and vast changes in the middle, and little and slow changes again at the end. For example, Chambers (2006) reported the shift of the past tense of "sneak" in Golden Horseshoe, Canada, the past tense of "sneak" is traditionally "sneaked", however "snuck" appeared and spread. Figure 1.1 shows the ratio of the people who use "snuck" according to the age groups plotted on the X axis.

Figure 1.1 clearly shows that the shift from "sneaked" to "snuck" is represented by an s-shape curve. S-shape curves are also known to resemble to and correspond to logistic regression functions as shown in Figure 1.2. In language studies, the X-axis corresponds to time, such as ages of the participants and the time of data collection. The degrees of language changes are plotted on the Y-axis, representing the replacement of an old form by a new form.

Inoue (2000), indeed, has investigated language changes over forty years in Japanese. In his study, time is defined as the physical time of data collection to be plotted on the X axis. The language changes to be plotted on the Y axis were represented by the standardization of a regional variety. The data were collected three times by the National Institute for Japanese Language and twice by Inoue (2000) in Yamagata, a northeast region of Japan, in which a distinct local dialect is often observed. His data demonstrated that the language shift was described by an S-shaped curve by plotting the ratio of the regional and the counterpart standardized forms observed in the given geographical area over the five times of data collection.

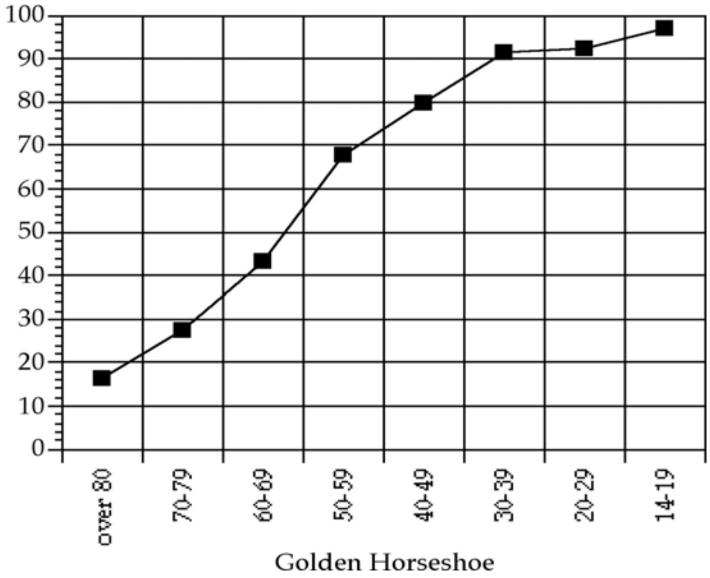


Figure 1.1. Change of past tense of “sneak”: “sneaked” vs. “snuck”

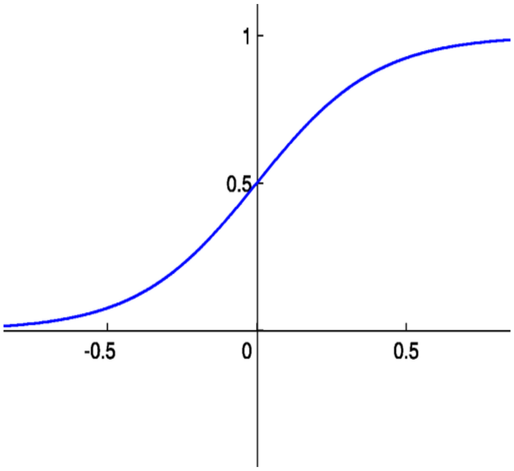


Figure 1.2. Logistic function

Studies, including but not limited to Inoue (2000), have shown empirical evidence as to the efficiency of the S-shaped curves to describe language changes, however, not much is available yet as to the theoretical and methodological research on the nature of the S-shaped curves and the mechanisms of language changes behind the observed phenomena. Thus, this paper provides a methodological suggestion for analyses of language changes with multiple variables by applying a logistic regression model. It also introduces a theoretical explanation for the mechanism behind language change phenomena, which can often be described by S-shaped curves, by employing a psychophysical model.

2. Models for language changes 1: Linear regression model

Altmann, Buttlar, Rott, and Strauss (1983) provided an approximate curve, an s-shape curve, which describes language changes. The model is expressed as follows:

$$p = 1 / \{1 + A\exp(-Kt)\}, \qquad (1)$$

where p stands for observed ratio of responses and t represents time. K and A are constants, which are calculated the by least square method. It should be noted that Altmann et al. (1983) made a significant contribution, theoretically and methodologically, by explicitly describing language changes with a mathematical model with time as the variable. In fact, a number of studies have been conducted based on their model to describe language changes over time, particularly in Europe. For example, Sanada (2002) described the decrease of Sino-Japanese words in dictionaries of technical terms.

Equation (1) can be transformed in Equation (2) as follows:

$$\log\{(1/p) - 1\} = \log A - Kt, \qquad (2)$$

which is, apparently, a logistic function. Logistic models, in fact, represent the "density effect" model in population and biology studies. The density effect refers to the phenomena, in which increase of the population is interfered when the increase exceeds the capacity of the given environments. For example, excessive increase of population might trigger spread of diseases or lack of food. In short, the model of language changes by Altmann et al. (1983) as expressed by Equation (1) is theoretically identical to the density effect model as expressed by Equation (2). Indeed, a simplest differential equation of density effect can be expanded to Equation (1) when the density of population is proportionately allocated, indicating a theoretical comparability between the two models. In other words, the comparability among the models indicates that the principle mechanism behind the phenomena of changes may be commonly shared across areas of studies.

The model of language changes by Altmann et al. (1983) should be noted, however, their model leaves a few issues unsolved. First, the model as expressed by Equation (1) does not allow computation when the value of p is zero or 1.0. For example, let us go back to the example of the past tense form “snuck” of “sneak”. Suppose that the new form “snuck” is so popular that all the participants in their 20s and 30s use it. The probability of “snuck” is 100%, i.e. the p value equals to $p = 1.0$, with 0% of the counterpart “sneaked”. Then, the left part of Equation (2) cannot be calculated because log 0 does not theoretically exist. Such a result consequently denies the existence of the given responses, i.e. “snuck”, by the participants of the younger population in this example, which contradicts with the actually observed phenomena. In other words, the model is invalid for prediction of a language shift from an old to a new form.

The second issue is concerned with the candidate variable included in the model. As shown in Equations (1) and (2), time is the only variable to describe language changes, which is obviously counter-intuitive. Language changes presumably involve multiple variables, including physical, social, and psychological factor. In addition, “time” can be represented by multiple measures, such as “age of the participants” and “time of data collection”. Thus, accurate description and prediction of language changes should consider multiple variables at the same time, including those related to time.

This paper proposes a logistic regression model by the maximum likelihood estimation as a solution for the two issues mentioned above. The model is expressed as follows:

$$\log\{p/(1-p)\} = Z\,, \tag{3}$$

where p refers to the probability of response ratio in a binary-option task and *log* is the logarithm to the base e. Z stands for responses of a binary-option task, which can be expressed by a linear function with multiple variables as in $Z = a_1x_1 + a_2x_2 + b$. This paper refers to $p/(1-p)$ as *odds* and $\log\{p/(1-p)\}$ as *logit*. Logistic regression models, in other words, are expressed with logits on the left side and multiple regression models on the right side in the equations. Equation (3) can be further transformed to Equation (4) as follows:

$$p = 1/\{1 + \exp(-Z\,)\}, \tag{4}$$

where *exp* stands for exponential function. Equations (3) and (4) can be expanded to Equations (5) and (6), with X_1 being time and a_1 and b being the constants.

$$\log\{\ p/(1-p)\ \} = a_1x_1 + b, \tag{5}$$

therefore

$$p = 1/\{1 + \exp[-(a_1x_1 + b\ )]\ \}, \tag{6}$$

which is a logistic regression model commonly applied to medical studies, exhibiting its efficiency to predict the ratio of occurrences in binary phenomena.

A logistic regression model as expressed by Equations (5) and (6), which is originated from Equation (3), is seemingly applicable to studies on language changes. In fact, Equation (3) can be deduced from Equation (2) by mathematical manipulation. Equation (2) can also be transformed in Equation (2.a.) as follows:

$$\log\{ p/(1-p) \} = Kt - \log A , \qquad (2a.)$$

where t represents time and A and K are constants.

Applications of logistic regression analyses are, indeed, found in sociolinguistic studies as well. Labov (1972), for example, employed a logistic regression model to analyze empirical data. Studies with logistic regression analyses are found in Japanese as well (Hibiya, 1988; Matsuda, 1993, Yokoyama & Wada, 2006), all inductively showing empirical applicability of logistic regression models to describe language changes. Little research, however, theoretically explained the mechanism behind the S-shaped curves and logistic regression models. Thus, it should be of contribution to theoretically explain the S-shaped curves of language changes. Such research should also provide a method of analyses of language changes. Thus, this paper proposes a method to analyze language changes, which can be represented by an S-shaped curve with multiple variables, by applying a multiple logistic regression models.

3. Models for language changes 2: Logistic regression models

Sociolinguistic studies have shown that language changes involve multiple variables, such as contexts, environments, genders, and geographical regions. Thus, it should be methodologically necessary to consider multiple variables in a model to express language changes, so the critical factors are identified among various candidates. It should also be noted that models should be valid and applicable to a wide range of data possibly observable, including but not limited to, probability values of zero and 1.0, which previous models could not accommodate. The first purpose of this paper, therefore, is to propose a model, a logistic regression model with multiple variables, in specific, as a methodological solution.

This section simulates the applicability of the proposed logistic regression model using existing data by Chambers (2006). For example, let us consider the psychological condition of the participants in the data described in Table 1. Suppose that the participants of 50s and 60s were in a marked psychological condition, due to the environment of data collection, for example, or to the incomepetence of the interviewers. Also suppose that those participants in a marked psychological condition used the new form "snuck" 30% to 50% more often than in normal psychological conditions. Table 1 below summarizes such a hypothetical data, with the two independent variables, i.e. age and psychological condition, and with the independent variable, i.e. ratios of observed occurrences of "sneaked" vs. "snuck". The value 1 for the psychological condition denotes a marked psychological condition, whereas zero represents a normal condition.

Table 1
Hypothetical data modified from Chambers (2006)

Age	Psychological condition	Ratios of observed occurrences of "snuck"
85	0	18
75	0	28
65	1	95
55	1	95
45	0	80
35	0	91
25	0	92
15	0	98

The values of parameters in the model were estimated based on Equation (7) as follows:

$$Z = a_1x_1 + a_2x_2 + b, \qquad (7)$$

where Z represents the ratio of "snuck" responses as opposed to "sneaked", x_1 and x_2 denote the age and the psychological condition, respectively. The values of the constants a_1 and a_2 were a_1= –0.076 and a_2 = 0.970, respectively, and the b value was b = 4.782 by the method of maximum likelihood estimation.

The probability estimated by the logistic regression model is expressed by Equation (6) as follows:

$$p = 1 / \{ 1+ \exp [– (a_1x_1 + a_2x_2 + b)]\}, \qquad (6)$$

as mentioned earlier. Thus, the rate of "snuck" occurrences against "sneaked" to be estimated is obtained by applying the values of constants to Equation (6), which is represented as follows:

$$p = 1 / \{ 1 + \exp(0.076\, x_1 - 0.970\, x_2 - 4.782)\}, \qquad (6.a.)$$

which yielded the predicted values of probability P as shown in Figure 3.1 below. Figure 3.1 also shows the observed occurrence rate of “snuck”. Likewise, Figure 3.2 shows the observed and predicted occurrence rates of “snuck” based on the regression model with single variable as expressed by Equation (1) by least square method. It is visually clear, based on Figures 3.1 and 3.2, that the multiple regression model as expressed by Equation (6) predicted the responses more precisely than the model with single-variable did.

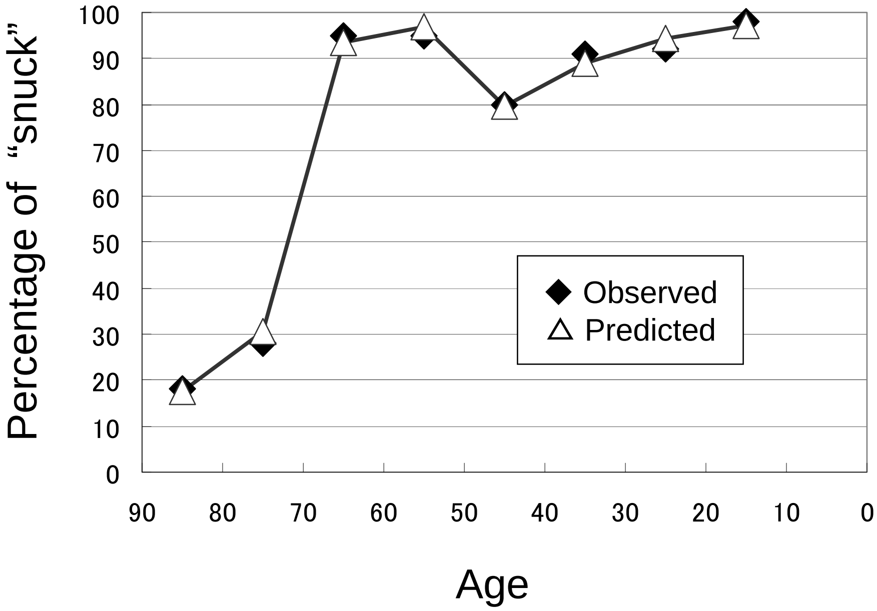


Figure 3.1. Prediction from multiple logistic regression model as expressed by Equation (6)

These results indicate that logistic regression models of the S-shaped curves allow analyses with multiple variables, which was not available with single-variable regression models without less significant variables. Analysis with multiple variables is apparently more powerful and accurate, as well as intuitively valid, and thus, more precise analyses and prediction should be available thorough multiple logistic regression models. For example, for sociolinguistic studies, genders, occupations, and locations of data collection can be included in the analysis together. For investigation of written languages, for example, writer-dependent variables, such as genders, and text-dependent variables, such as word types and genres, can be included in the analyses at the same time.

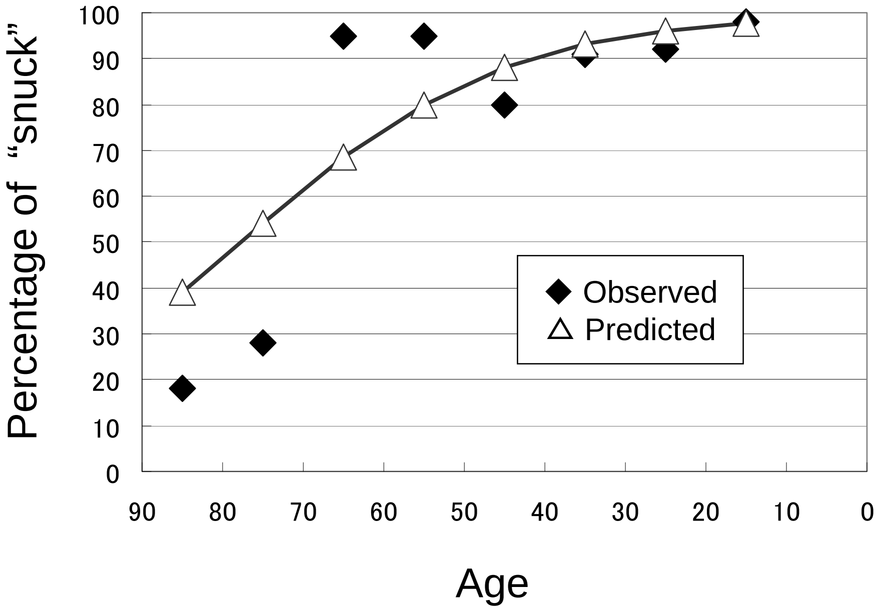


Figure 3.2. Prediction from single logistic regression model

4. Psychophysical model and the S-shaped curve of language changes

The second issue, which this paper addresses, is the theoretical explanation for a mechanism behind the phenomena of language changes. Previous research has addressed this issue, of course. For example, Altman et al. (1983) employed a model to describe language changes, which is represented by S-shaped curves. Their model was, in principle, a differential function of the density effect. Thus, their model introduced that the density effect affects language changes, though implicitly, with the density being a factor external of language users.

This paper, in contrast, asserts that a psychological behavior, which is internal of language users, reflect languages changes, consequently yielding S-shaped curves. It is assumed that language changes are produced by binary options in human psychology, which is based on relative strength of the two competing forms in the memory of language users (Yokoyama, 2006, 2007). It should be noted here, intuitively, empirically, and experientially that behavior and changes involving languages are attributable to both user-internal and external variables, possibly with interactions of the two as well. However, for the purpose of argument in this paper, a psychological factor, i.e. user-internal variable, is a major issue to be discussed. In specific, the discussion focuses on the role of memory strength, which is attributable to relative frequencies of the two competing forms of languages.

4.1. Logistic regression models and normal distribution

Regression models often assume normal distributions, and a part of the area is a critical measure used in statistical analyses. The cumulative sum of area, based on a normal distribution, is represented by an S-shaped curve as in Figure 1.2. as pointed out in testing research, an area of studies in psychological testing. The S-shaped curve can be expressed by a model as follows:

$$p = 1 / \{1 + \exp[-Da(\theta - b)]\}, \qquad (7.1)$$

where p is the accuracy rate of the test, b represents the degree of difficulty of the test, a does the distinctive power of the test, and θ is a constant. The b is an inflection point as well. Equation 7.1 above can also be expressed as follows:

$$p = 1 / \{1 + \exp(-DZ)\}, \qquad (7.2)$$

when $a = K$, $\theta = t$, $-ab = B$ and $Z = Kt + B$. Equation 7.2 is an approximation of the cumulative distribution function of a normal distribution if the value of D equals to $D = 1.7$, according to Lord and Novick (1968). Such comparability between Equation (7.2) and the cumulative distribution function of a normal distribution allows comparability between Equation (7.2) and a logistic regression model.

Equation (7.2), indeed, can be expanded to a logistic regression model if the value of D equals to $D = 1.0$ as expressed by Equation (4). Such comparability indicates that logistic regression models are an approximate representation of the cumulative distribution function of a normal distribution. In order to verify whether logistic regression models efficiently represent the cumulative distribution function, the differences were computed between the cumulative distribution function of a normal distribution with the standard deviation of 1.7 (SD = 1.7) and the values obtained from the logistic regression model, as shown in Figure 4.1.1. The errors fall in between ±0.01, indicating that logistic regression models can represent the cumulative distribution function with SD = 1.7 quite precisely, as well-known in testing research.

Cumulative sum of areas of a normal distribution can be mathematically obtained by integral calculation, however, the calculation is complex and demanding. On the contrary, logistic regression models can be computed without mathematical manipulations, and are thus an efficient method to approximately represent the cumulative sum of areas of a normal distribution.

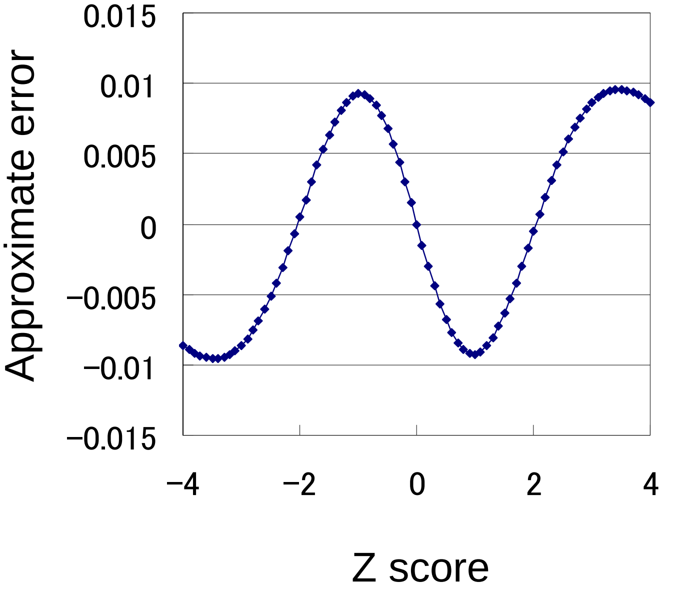
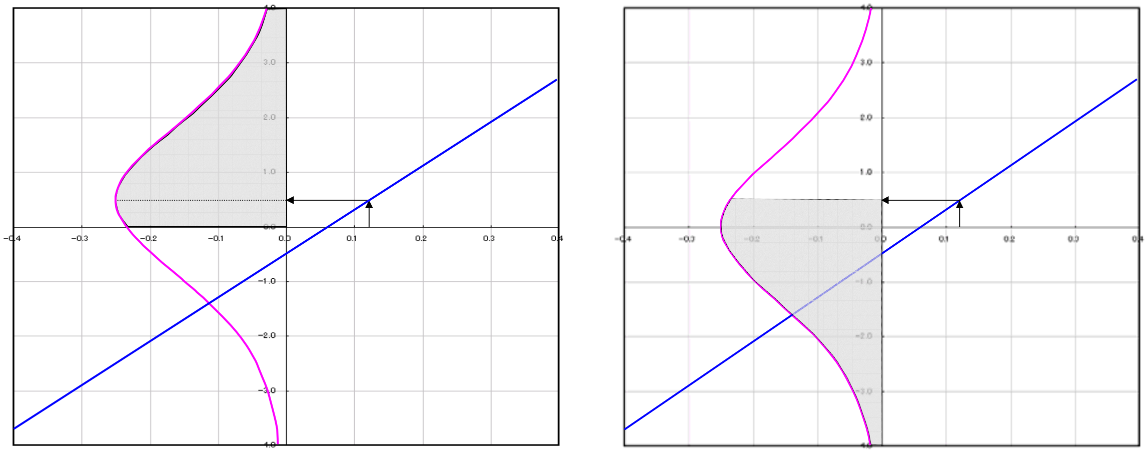


Figure 4.1.1 Differences between the cumulative distribution function of a normal distribution with the standard deviation of 1.7 (SD = 1.7) and the logistic regression model

4.2. Threshold model: Success, failure, and instability

Methodological efficiency of logistic regression models to represent a normal distribution also allows the computation of probability by a threshold model. For example, let us cite an example of language changes in Canada, in which the past tense form of “sneak” exhibits an old and a new forms, i.e. “sneaked” and “snuck”. Suppose that the use of new form “snuck” is denoted by a “success” for the sake of methodological efficiency, following the convention which encodes a set of binary options as either “success” or “failure”. Also suppose that variable *Z* represents the competency that enables “success” and that the variable Z is determined by variable *x*, i.e. age of the participants, in the “sneaked” vs. “snuck” example. Logistic regression assumes that the participants sharing the same *Z* values may produce either “success” or “failure”. In other words, “success” or “failure” is a probabilistic phenomenon, which allows both “success” and “failure”. The resultant phenomena, i.e. either “success” or failure”, depends on variable *Z*, which is dependent on variable *x*, but not a direct and linear function of *Z*. The probability for one of the either form to be observed is the highlighted area of the normal distribution represented by Figure 4.2.1. The *x* axis indicates age and the *Y* does the *Z* values, which can be obtained from the linear function. “Success” or “failure”, however, may vary unstably, depending on other variables besides *Z* values of competency, such as contexts, physical and psychological conditions, and luck. Such instability may vary, either positively or negatively, contributing to the “success”, however, its variability is assumed to exhibit regularity with a normal distribution with the *Z* value at the median in psychophysics. For example, a “success” reflects competencies as well as luck, and neither of the two alone guarantees a “success”. In other words, the normal distribution of the variability represents a combination of competency and instability.



(a) (b)

Figure 4.2.1 Sum of the area above the threshold equals to the probability value *p*

Let us suppose that the normal distribution curve in Figure 4.2.1(a) moves along the *Y* axis, i.e. the *Z* values. Suppose that the distribution curve stops at the point, at which the mean of the normal distribution coincides with the *Z* value corresponding to a given age, i.e. variable *x*. The standard deviation value of such a distribution is SD = 1.7. The point at which $Z = 0$ is called a threshold, and the sum of the area above the threshold equals to the probability value p, i.e. rate of the new form “snuck” to be chosen against the old form “sneaked”.

The probability value p equals to that obtained by a procedure as follows. First, imagine moving the normal distribution curve in Figure 4.2.1(b) along the Y axis, i.e. the Z values. Second, move the distribution curve until its mean of the normal distribution falls where Z value equals to zero. In other words, the mean, i.e. the peak of the distribution curve, falls at $Z = 0$ in Figure 4.2.1(b), whereas the peak is at any of the given values of Z in Figure 4.2.1(a). The value of the standard deviation equals between the two curves represented by Figures 4.2.1(a) and (b), with SD = 1.7. Third, define the threshold according to the given value of variable x, i.e. age. Fourthly, compute the sum of the area below the threshold. The value of the sum of areas computed as this is the probability value p for “success”.

The method described above is applicable to categorical data as Table 3.1, allowing us to estimate that “snuck” will be observed if the probability value is $p > 0.50$. Likewise, “sneaked” is expected when the probability value is $p < 0.50$. Furthermore, this model can handle multiple variables, which is not explicitly treated in the statistical references known to the authors.

The threshold model is based on normal distributions, however, a logistic regression model is also applicable to computation of the probability values p, because logistic regression models can efficiently represent the cumulative sum of areas of a normal distribution with a standard deviation of 1.7. The logistic regression model to be used for such a purpose as probability computation can be expressed as follows:

$$p = 1 / \{1 + \exp(-Z)\}, \tag{4}$$

4.3. Exposure relativity theory: Memory strength and S-shaped curves

The purpose of this section is to present a model of Z values and to provide an explanation as to why the threshold model yields varying values for Z.

Let us assume, to start with, that a shift from the old form “sneaked” to the new form “snuck” reflects the relative frequency of exposure to the two competing forms. In other words, the more you are exposed to the new form, the more likely you choose the new form against the old counterpart. Such a role of relative frequency has been, indeed, introduced as the exposure relativity theory by Yokoyama (2006), based on the mere exposure effect and the generalized matching law.

Exposure relativity theory asserts as follows. As shown by the studies of the mere exposure effect, repeated exposure to unfamiliar forms of languages increases favorability for the new forms. Increase of favorability contributes to more frequent use of the new form in the given community, which, in turn, relatively decreases the frequency of the old form. Increased use of the new form increases frequency of exposure to the new form even more, whereas less use of the old form decreases the relative frequency of exposure to the old form in the given community. Such changes of relative frequencies consequently reinforce the use of the new form even more. In other words, the exposure relativity theory asserts that use and exposure reciprocally reflect each other. The model of exposure relativity theory (Yokoyama, 2006) is expressed as follows:

$$\log \{p/(1-p)\} = S \log (r_1/r_2) + \log b\,, \qquad (8)$$

where p represents the frequency of the new form "snuck", r_1 the frequency of exposure to the new form "snuck", r_2 the frequency of exposure to the old form "sneaked", with S being sensitivity and log b being a response bias. Equation (8) is a logistic regression model, with the right side log (r_1 / r_2) representing the relative frequency of exposure called "exposure relativity" (Yokoyama, 2006) that contributes to the preference for the new form "snuck" against the old form "sneaked". Equation (8), in fact, can be transformed to (4), i.e. a logistic regression model with multiple variables as well. The discussion of this paper employs a logistic regression model as a practically efficient representation of the threshold model based on a probabilistic perspective, which assumes a normal distribution.

The model expressed by Equation (8) can explain a wide range of variations, such as regional and generation variations. For the sake of argument, let us take variations according to ages as an example. It is conceivable that older people are exposed to the old form "sneaked" more frequently than to the new form "snuck". Thus, let us assume that the relationship of exposure frequencies between the new form "snuck" and the old counterpart "sneaked" is expressed as $r_1 < r_2$, where r_1 denotes exposure frequency to the new form and r_2 refers to that of the counterpart old form. Let us also assume that the younger generation is exposed equally to the new and old forms, i.e. $r_1 = r_2$. These assumptions naturally lead to more frequent observations of the new form "snuck" among the younger generation than among the older generation, as represented by an S-shaped curve as in Figure 1.1, for $r_1 < r_2$ applies to the older generations whereas $r_1 = r_2$ does to the younger generations. The Z values can be obtained by Equation (9) as follows:

$$Z = S \log (r_1/r_2) + \log b\,, \qquad (9)$$

by simultaneously equating Equations (3) and (8). In sum, Equation (9) is a proposed model of language changes from one form to the other, i.e. a binary option, obtained by mathematical manipulations. Therefore, further validation with empirical evidences is necessary. Further theoretical research is also called for because the model is a deductive explanation of the given phenomena, i.e. *Z*, and therefore it is an expression of the mechanism of phenomena.

The theoretical explanation proposed here, which is based on the exposure relativity theory, expresses the effect of relative frequencies of exposure in human memory, in that more frequent items provide relatively more strengths in memory than less frequent items do. For example, old generations are more frequently exposed to the old form "sneaked" compared to the new form "snuck", therefore the relative strength of the old form "sneaked" is stronger in their memory compared to the counterpart new form "snuck". Such a relative gap of strength in memory consequently leads to more probability to choose the old form and less probability to choose the new form. On the other hand, the younger generations are, presumably, exposed to the old form "sneaked" less frequently than the older generations due to their limited experience with the old form in their environments. Because the younger generation has not got much exposure to the old form, the relative exposure frequency to the new form is higher among the younger generations than among older generations.

For example, let us suppose that the old form "sneaked" was used once yesterday. The older generation has seen it yesterday as well as many times over the past tens of years of their lives. Thus, the total number of times, for which they have seen it could be ten times, hundred times, etc. cumulatively. On the other hand, suppose that the younger generation has seen the old form "sneaked" only yesterday but never before for they haven't lived long enough to get many opportunities to see the old form. Thus, the younger generation has seen the old form "sneaked" only once in their lives. Suppose the counterpart new form "snuck" was also used once yesterday, which results in the same frequency of exposure for both the older and younger generations as for yesterday. Then, the relative frequency of exposure to the new and old forms, i.e. r_1 / r_2, for the older generation is $r_1 / r_2 = 1 / 10 = 0.1$, whereas that for the younger generation is $r_1 / r_2 = 1 / 1 = 1.0$. Therefore, the new form "snuck" leaves more room and strength in the memory of the younger generation, whereas it only takes up a very small impression, or strength, in the memory of the older generation in their entire lives. Such a relatively smaller strength of the new form leads to an only small probability to choose the given new form among the older generation. Attention to memory, i.e. a mechanisms that is internal of language users, and the assumption of frequency effect, i.e. user-external factor, is presumably critical to research on language changes, because language use is, in principle, a social behavior produced by human minds. By the same token, recent brain studies, which focus on user-internal mechanisms, should contribute to research of language changes and behavior, provided that such studies consider social and inter-personal interactions that intra-personal factors experience.

A shift from an old form to a new one is comparable to the standardization of regional varieties. For example, Inoue (2000) asserts that the speed of standardization of a regional variety reflects the amount of inter-personal contacts, i.e. geographical neighborhood effect, and frequency of the observed phenomenon of the given language shift. He further proposes a model of standardization speed of regional varieties as follows:

$$V = f\,(\,C \cdot freq\,), \qquad \text{Modified from Inoue (2000)}$$

where V denotes the speed of language shift, C represents the amount of inter-personal contacts, and *freq* does the frequency of the observed phenomenon of the given language change with f being a given function. The variables in his model are inter-personal contacts and frequency of language phenomenon, to which exposure frequency to language forms is attributable, in the given communities. His model thus consequently supports the basic principle/concept of the exposure relativity theory which considers exposure frequency as a major variable explaining language phenomena.

Little research is available so far, which considers probability of language change phenomena to be observed. Conceptual connection between the multiple logistic regression model and the exposure relativity theory should contribute to the advancement of the language studies, for the model expresses the S-shaped curve of language changes, as discussed in this paper, and the theory accounts for a basis which interweaves both psychology and observable phenomena, i.e. intra- and inter-personal variables, both of which are presumably critical to language changes.

5. Summary and discussion

This paper first demonstrated the efficiency of the S-shaped curves by analyzing a hypothetical data with a logistic regression model. In addition, estimation and probabilistic use of the models were not explicitly introduced in research on language changes.

The paper further proposed a theoretical view as to the mechanism behind the S-shaped curves of language changes. Altmann and his associates proposed the notion of S-shaped curves to describe language changes, however, their theoretical background is based on socio-linguistic perspective, focusing on inter-personal factors observable in the given communities. This paper, in contrast, employs a psychological perspective, which considers intra-personal variables as a critical factor. In other words, this study and the series of studies by Yokoyama (2006, 2007) considers memory to be precise, i.e. an intra-personal variable, as a critical factor governing the mechanism of the S-shaped curve of language change phenomena. Such a psychological perspective has a long tradition in psychology, which focuses on intra-personal variables, however, the model proposed in this paper includes both social phenomena, i.e. inter-personal, as well as memory, i.e. an intra-personal psychological mechanism. In addition, it argues that both of the two factors reciprocally affect each other and play critical roles in the mechanism of the S-shaped curves to explain the language changes. This paper further applied the proposed model to the hypothetical data in this paper for the purpose of prediction. The proposed model should contributes to the research of language changes by allowing researchers to include and combine the two distinct variables, i.e. inter-personal and intra-personal ones, interwoven in a quite simple and efficient model.

The model should allow identification of the most critical factors that explain the language changes over a long period of time, when the nation has experienced vast economic and socio-cultural changes.

References

Aitchison, J. (1991) *Language change: progress or decay?* 2nd ed. Cambridge: Cambridge University Press.

Altmann, G., von Buttlar, H., Rott, W., Strauss, U. (1983). A law of change in language. In: Brainerd, B. (ed.), *Historical Linguistics: 104-115* (=*Quantitative Linguistics.* Vol. 18). Bochum: Brockmeyer..

Baum, W.M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior 22, 231-242*

Belke T.W., Belliveau J. (2001). The general matching law describes choice on concurrent variable-interval schedules of wheel-running reinforcement. *Journal of the Experimental Analysis of Behavior 75, 299 - 310.*

Chambers, J.K. (1998). Social embedding of changes in progress. *Journal of English Linguistics 26, 5-36.*

Chambers, J.K. (November, 2006). Paper presented at the meeting of National Institute for Japanese Language, Tokyo, Japan.

Collett, D. (2003). *Modelling Binary Data*. London:Chapman and Hall.

Elliot. R., Dolan. R. (1998). Neural response during preference and memory judgments for subliminally presented stimuli: A functional neuroimaging study. *The Journal of Neuroscience 18, 4697-4704.*

Fagen, R. (1987). A generalized habitat matching rule. *Evolutionary Ecology 1, 5-10*

Hibiya, J. (1988). *A quantitative study of Tokyo Japanese.* Doctoral Dissertation, University of Pennsylvania, 1988.

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

Lord, F., Novick, M. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley.

Matsuda, K. (1993). Dissecting analogical leveling quantitatively: The case of the innovative potential suffix in Tokyo Japanese. *Language Variation and Change 5, 1-34.*

Moin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology 85, 1035-1048*.

Sanada, H. (2002). *Kindai nihon-go ni okeru gakujutsu-yougo no seiritsu to tei-chaku* [Emergence and establishment of academic terminologies in modern Japanese]. Tokyo, Japan: Junbunsha.

Woolverton, W.L, Alling, K. (1999). Choice under concurrent VI schedules: comparison of behavior maintained by cocaine or food. *Psychopharmacology 141, 47-56.*

Yokoyama, S. (2006). Mere exposure effect and generalized matching law for preference of Kanji form. *Mathematical Linguistics 25, 199 - 214.*

Yokoyama, S., Wada, Y. (2006). A logistic regression model of variant preference in Japanese kanji: an integration of mere exposure effect and the generalized matching law. *Glottometrics 12, 63 -74*.

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology 9, 1-27.*