Glottometrics 12, 2006, 63-74

A logistic regression model of variant preference in Japanese kanji: an integration of mere exposure effect and the generalized matching law

Shoichi Yokoyama, Yukiko Wada[1]
The National Institute for Japanese Language, Tokyo

Abstract: The word *hinoki* or 'cypress' can be transcribed in two variant forms, 檜 (the so-called "traditional" variant) and 桧 (the "simplified" variant), in Japanese kanji. Such variant forms are called *kanji variants*. The present paper reviews a series of studies on Japanese kanji recognition (Yokoyama, 2006a, 2006b, 2006c), and proposes a model which accounts for performance in a preference judgment task based on kanji frequency data. Yokoyama (2006a) administers preference judgment task in which the participants were presented with 263 pairs of traditional and simplified variants and asked to choose the more preferable variant of each pair. The analyses indicate a positive contribution of frequency to variant preferences, supporting the so-called "mere exposure effect" theory of Zajonc (1968). This finding leads to a logistic regression model that describes preference behavior in kanji recognition, based on Fechner's law. Yokoyama (2006b) shows that the model is comparable to the so-called "the generalized matching law" of Baum (1974) and to "the ideal free distribution theory" of Fagen (1987). Yokoyama (2006c) further examines the predictive validity of the model with empirical data obtained from a preference judgment task, administered in the Tokyo and Kyoto areas. Logistic regression analyses are performed with the ratio of preference for the given variants and the logit of the character frequencies, yielding significant correlations between the predicted probabilities and the observed responses (r = .804 for Asahi newspaper data). The present paper synthesizes these studies and proposes a logistic regression model that efficiently describes preference behavior in Japanese kanji recognition, integrating the theoretical perspectives of mere exposure effect and the generalized matching law.

Key words: mere exposure effect, Fechner's law, generalized matching law, logistic regression analysis, variation theory, kanji, variants, preference, familiarity

1. Variant Forms in Japanese Kanji Characters

1.1. Kanji Variants in Japanese Language

Pairs of kanji characters that share the same meaning and pronunciation but exhibit varieties in their visual forms are called *variants*. Variants are commonly found in Japanese kanji characters. For example, the pair of kanji "桧" and "檜", both representing "cypress", has two variants: the traditional "檜" and the simplified "桧". Very few studies have provided insights into familiarity of and preference for orthographic variants with inter-disciplinary perspectives (Sasahara & Yokoyama, 2000). Yokoyama (2006a) employs a two-alternative

[1] Address correspondence to: Shoichi Yokoyama or Yukiko Wada,
The National Institute for Japanese Language, Midori-cho, Tachikawa-shi, Tokyo, 190-8561, Japan.
E-mail: yokoyama@kokken.go.jp or ywada@kokken.go.jp.

forced-choice task, in which the participants are presented with pairs of kanji variants and asked to choose the item in each pair that is more preferable. The participants were instructed to perform the tasks assuming that they were word-processing with digital tools, such as computers or cell phones. The number of participants was approximately 200; they were college students in the Tokyo area. The data indicates that the participants' preference for variants are not attributable to graphic complexity or historical reasons, but to the frequencies of the given variants. Yokoyama (2006b) further analyzes the contribution of the frequency data, introducing a logistic regression model that accounts for performance in the preference judgment task based on frequency data obtained from a newspaper corpus.

1.2. Scope of this paper

The purpose of the present paper is to review related studies and to propose a logistic regression model that accounts for preference performance in kanji character recognition by integrating mere exposure effect and the generalized matching law. Yokoyama (2006a) demonstrates that character frequency data from Asahi and Yomiuri Newspaper corpora can account for performance in a preference judgment task in which 263 pairs of variants were presented to native speakers of Japanese. A subsequent study by Yokoyama (2006b) reveals the corresponding relationship between the logistic regression model, introduced by Labov (1972) in sociolinguistics, and the generalized matching law, commonly applied in animal behavioral research. Yokoyama (2006c) further proposes a quantitative model that accounts for performance in the preference judgment task, which is reviewed in this paper. Such a model is expected to provide an innovative framework to investigate recognition and use of orthographic variation, because linguistic activities involve choices among multiple alternatives, and because such choices are often based on preferences for certain forms.

1.3. Mere Exposure Effect in Social Psychology

A number of studies have shown that the preference mechanisms in human psychology involve memory and cognitive factors. Zajonc (1968), for example, has proposed the so-called "mere exposure effect" in social psychology. Based on the positive correlation between frequency and preference, in which high-frequency words are more preferred than low-frequency words, he asserts that repeated exposure to unfamiliar items increases preference frequency is a contributing factor for preference development (Zajonc, 1968).

Following the framework of mere exposure effect, various studies have shown that the mere exposure effect is observed regardless of the participants' awareness of the exposure. A method used by studies in brain research, among others, is to present kanji characters as primes and then examine whether the primed stimuli are preferred, in comparison to un-primed items. In Elliot & Dolan (1998), each of the twenty primes was presented for .05 second, followed by a masking stimulus for .45 second, ten times in all. The total time used to present the primes was 1 minute and 40 seconds. The participants were native speakers of English, completely unfamiliar with kanji. Thus, it was assumed, in the given condition, that the participants were not able to recognize the prime characters, and that they were even unable to perceive the visual representations of the kanji characters. However, the results showed that they preferred the primed stimuli to the un-primed counterparts, indicating the contribution of primes, i.e. exposure, to preference judgment performance.

1.4. Frequency, Preference, and Familiarity

Monin (2003) has shown that preference for certain items increases perceived familiarity, whether the exposure was conscious or unconscious. In his experiments, the participants judged their familiarity to stimuli that exhibited varying degrees of positive attributes. For example, the stimuli included photographs of beautiful-looking persons and real words that were semantically positive. The participants' task was to judge their perceived familiarity to such stimuli. Monin's analyses have shown that positive attributes of the stimuli, such as physical and semantic characteristics, contributed to performance in the familiarity inference task. In other words, items with positively preferred attributes were perceived as being more familiar.

The findings of these studies indicate that exposure frequency increases familiarity and that familiarity and preference are closely related to each other. Therefore, the present study assumes that exposure to kanji, which is operationally defined as frequency, contributes to preference, which is defined as performance in the preference judgment task in the study. It is also assumed that preference and familiarity are closely related, reflecting each other.

2. Fechner's Law and Psychophysical Model

2.1. A Model to Predict Familiarity Judgment Performance Based on Frequency

Yokoyama (2006a) proposes a model to account for performance in a variant preference task based on Fechner's law, which is often applied in perception studies in psychophysics. Fechner's law defines the perception degree as *S*, strength of stimulus as *I*, common logarithm as *log*, slope as constant *K*, and the intercept on the *S* axis as constant *C* as follows:

$$S = K \log I + C, \tag{1}$$

where Equation (1) is a linear combination equation. By analogy with Equation (1) above, kanji familiarity is expressed as follows:

$$Familiarity = K \log (frequency) + C, \tag{2}$$

Based on Equation (2), the familiarity of traditional variants, referred to as *FamTrad* hereafter, is expressed as follows:

$$FamTrad = K \log (FreqTrad) + C, \tag{3.1}$$

where the frequency of the given traditional variants in the newspaper corpora is defined as the *frequency of traditional variants* (*FreqTrad* hereafter), slope as *K*, and intercept on the *S* axis as *C*.

In the same way, the familiarity of simplified variants, referred to as *FamSimp* hereafter, is expressed as follows:

$$FamSimp = K \log (FreqSimp) + C, \tag{3.2}$$

where *FreqSimp* stands for the frequency of the simplified variants.

A value of 1 is added to the frequency data in advance to avoid zero values in logarithm computation. Namely, the frequency of a traditional variant is 1 plus the frequency of that

traditional variant, the frequency that of a simplified variant is 1 plus the frequency of that simplified variant, respectively, in the subsequent analyses. This conversion method is commonly employed in various studies.

2.2. A Model to Predict Preference Performance Based on Character Frequency

As discussed in Yokoyama (2006a, 2006b), it can be assumed that preference for variants is attributable to frequency differences between the given pair of variants. Preference for traditional variants, referred to as *PrefTrad*, can be described with Equation (4.1) as follows:

$$PrefTrad = a\,(FamTrad - FamSimp) + b, \tag{4.1}$$

where the explanatory variable is defined as the difference in familiarities between the traditional and simplified variants, the slope as a, and intercept as b.

When Equation (3.1) is substituted for the familiarity of traditional variants in Equation (4.1), and Equation (3.2) for the familiarity of simplified variants in Equation (4.1), Equation (4.2) as the following is available:

$$PrefTrad = a\,K\,\{\log(FreqTrad) - \log(FreqSimp)\} + b, \tag{4.2}$$

Equation (4.3) below expresses preference for traditional variants as follows:

$$PrefTrad = a\,K \log(FreqTrad\,/\,FreqSimp) + b, \tag{4.3}$$

where the frequencies of the variants are transformed to a logarithm of the frequency ratio. This model may be extended to multiple regression models with other plausible variables, such as stroke numbers. However, this study will mainly examine a simple regression model, namely Equation (4.3).

3. Empirical Validation of the Simple Regression Model

3.1. The Simple Regression Model and Empirical Evidence

This section empirically tests Equation (4.3), which is a simple regression model introduced in the previous section. The model, in theory, describes performance in a preference judgment task on character recognition in Japanese kanji.

3.2. Variant Preference Task

3.2.1. Materials

Yokoyama (2006a) selected stimuli from 263 pairs of traditional and simplified variants in the CD-ROM data of Sasahara & Yokoyama (2000). The original 263 pairs of variants were selected based on the number of variants that each character exhibits. Technical issues were also considered, so that the stimuli could be displayed and printed with the 83JIS standard, which is the 1983 version of the Japanese Industrial Standards. Of the original 263 pairs, 86 pairs of variants were selected for the purpose of the study. Figure 1 shows some sample stimuli used in the task. Presentation order was randomized in the task.

Figure 1. Sample stimuli in the preference judgment task

3.2.2. Task, instruction, and procedure

The task was a two-alternative forced-choice task in a paper-and-pencil format. The cover page of the task asked about demographic information and experience in word processing with digital tools. The task was administered as a part of instruction in classes at colleges.

The participants were asked to suppose that they were word-processing and to choose the variant of each pair, i.e. either traditional or simplified, that they would prefer (Yokoyama, 2006a, 2006b). Word processing was chosen as the context of the task in order to minimize the effects of non-target variables. Effects of economy, such as efficiency in hand-writing, for example, which may lead the participants to prefer graphically simpler forms, were presumably minimized in the task.

3.2.3. Participants

The data were obtained from two groups of participants, the Tokyo and Kyoto groups, both of which consisted of college students. Data used in the analyses were obtained from the participants who met the following criteria: (1) native speakers of Japanese; (2) relatively younger generation, i.e. 25 years old or younger; and (3) have experience in word processing and typing Japanese characters. The Tokyo group consisted of eighty-five female college students in the Tokyo area, who participated in the study in the academic year 1996-1997. The Kyoto group was male and female college students in the Kyoto area, twenty males and fifty-two females, seventy-two in total, who participated in the study in 1998.

3.3. Frequency Data from Newspaper Corpus

Frequency data from the Asahi Newspaper were obtained from two sources: (1) Chikamatsu, Yokoyama, Nozaki, Long, & Fukuda (2000); and (2) Yokoyama, Sasahara, Nozaki, & Long (1998). The frequency data were based on a corpus of the Asahi Newspaper (Asahi--Shinbun-Sha, 1994), which included the text of the morning and evening papers from January 1 through December 31, 1993. Based on this corpus, the researchers manually corrected the inconsistencies between the hard-copy representations and the digital data. The total number of the kanji characters included in the data, i.e. the number of tokens, was 17,117,320 and the number of types was 4,546.

3.4. Analyses and Results

Preference for traditional variants, in percentage form, was computed for the 86 pairs of variants according to Equation (4.3) (discussed above), which is a simple regression model with logarithms of frequencies. Table 1 summarizes some of the computation results based on the Tokyo group data along with the frequency data based on the Asahi Newspaper corpus.

Table 1
Preference for traditional variants (PrefTrad) and frequencies in Asahi Newspaper

Pair	PrefTrad %	Frequency		Pair	PrefTrad %	Frequency	
		Simplified	Traditional			Simplified	Traditional
亜亞	2.4	1035	5	蝿蠅	10.6	11	1
壷壺	74.1	59	20	竈竈	12.9	1	0
陥陷	8.2	1285	0	条條	24.1	9948	116
奥奧	1.2	2590	0	嬢孃	12.9	108	0
蛍螢	54.1	98	2	飲飮	1.2	2274	0
学學	7.1	54725	7	真眞	15.3	12248	149
誉譽	3.5	2198	0	慎愼	15.3	2724	2
鴬鶯	65.9	16	4	槙槇	44.7	230	2
鴬鷺	25.9	16	0	靭靱	32.9	17	2
会會	4.7	161051	7	尽盡	2.4	1175	2
桧檜	71.8	230	15	侭儘	30.6	0	0
覚覺	1.2	4990	3	数數	1.2	29439	2
撹攪	10.7	12	2	薮藪	40.0	81	27
観觀	0.0	7794	0	銭錢	9.4	2503	1
潅灌	84.7	11	2	賎賤	65.9	2	9
狭狹	17.6	948	0	曽曾	20.0	926	13
尭堯	31.8	72	45	騒騷	3.5	1619	0
焼燒	3.5	2553	1	捜搜	5.9	5404	0
区區	1.2	28396	0	沢澤	38.8	14489	643
欧歐	24.7	8001	0	駅驛	10.6	3315	1
経經	5.9	38698	9	訳譯	3.5	3161	1
頚頸	81.2	5	40	鈬鐸	45.9	0	38
倹儉	7.1	36	0	単單	1.2	6886	0

A regression analysis was performed with Equation (4.3) with the frequency data from Asahi Newspaper as follows:

$$PrefTrad = 10.30 \log (FreqTrad / FreqSimp) + 41.88, \quad (4.3a)$$

where the explanatory variable was defined as the logarithms of the frequency ratio of the two variants.

The results show a significant correlation between the logarithm of frequency ratio and the preference for traditional variants with $r = .73$ ($p < .01$, $df = 84$), accounting for 52.90% of the variance. However, this analytic method may produce values for the estimate probabilities that are negative, 100, or greater than 100, suggesting that it is statistically invalid. A solution for this problem is introduced in section 4.2, in which logistic regression analysis is discussed.

The same procedure was applied to Equation (5.1) in order to compare the predictive power of Equations (4.3) and (5.1) as follows:

$$PrefTrad = a\ K\ (FreqTrad - FreqSimp) + b, \quad (5.1)$$

where the frequencies are not transformed to logarithms. Equation (5.1) can be expanded to Equation (5.2) as follows:

$$PrefTrad = 22.31\ (FreqTrad - FreqSimp) + 0.00, \quad (5.2)$$

which yielded little correlation, with $r = .21$ ($p < .05$, $df = 84$) in fact, accounting only for 4.61% of the variance. As Equation (5.2) differs from Equation (4.3) in that the frequency data are not transformed into logarithms, this result indicates that the logarithm of frequencies has stronger predictive power than the frequency data *per se*.

4. Application of the Matching Law in Mathematical Linguistics

4.1. The Generalized Matching Law in Animal Psychology

Yokoyama (2006b) demonstrated that kanji frequency data from the Asahi newspaper corpus explained the performance in the preference task by applying the generalized matching law. A constructive extension of this line of research is to evaluate the theoretical contribution of the generalized matching law and the role of frequency in written language.

The basic principle of the matching law was proposed by Herrnstein (1961), which was initially applied to behavioral studies of animals, and was expressed as follows:

$$R1 / (R1 + R2) = r1 / (r1 + r2), \quad (6)$$

where R refers to responses and r refers to frequencies of reinforcers. The mathematical model of the generalized matching law developed by Baum (1974), which expresses the relationship between the reinforcers and the response allocation as follows:

$$(R1/R2) = B\ (r1/r2)^{S}. \quad (7)$$

Equation (7) can be expanded to Equation (8) as follows:

$$\log (R1/R2) = S \log (r1/r2) + \log B, \quad (8)$$

where *log* refers to natural logarithms with base *e*, parameter *S* to the slope of the line representing sensitivity of reaction, and *log B* to the intercept representing response bias. Equation (8) is a logit equation, predicting the logit of response based on stimulus logit.

Ratio of reinforcement, i.e. *r1/r2* in Equation (8), and that of response allocation, i.e. *R1/R2*, may be explained by reference to an example experiment. Suppose that pigeons or rats in cages are rewarded with food by pushing levers called Levers 1 and 2, for example. The frequency of reward obtained by pushing Lever 1 is represented by *r1*, and that by pushing Lever 2 is represented by *r2*. The ratio of the frequencies of these two reward opportunities is referred as the ratio of reinforcers, i.e. *r1/r2*. The frequencies of lever-pushing behavior by the animals are represented as *R1* and *R2*, with *R1* referring to the frequency of the subjects' pushing Lever 1 and *R2* to that of pushing Lever 2. The ratio of these two frequencies, i.e. *R1/R2*, is referred to as the *ratio of response allocation*. Previous research in animal behavior has shown that response allocation is well expressed by Equation (8).

The generalized matching law seems to exhibit a wide range of applicability. In fact, it is comparable to the ideal free distribution theory in ecological studies, which describes the distribution of wild animal communities across multiple food sites (Fagen, 1987; Yamaguchi & Ito, 2006). The model of the ideal free distribution theory is identical to Equation 8, when the distribution ratio of individuals is replaced with *R1/R2* and the amount of food with *r1/r2*. More generally, it should be noted that the generalized matching law exhibits generalizability with empirical evidence across different fields of study.

Quantitative models, which are comparable to the generalized matching law, are applied in linguistic investigation as well. Well-known examples are sociolinguistic studies by Labov, in which linguistic variation and change are described by a logistic regression model (Wardhaugh, 1986). Labov's (1972) quantitative perspective is prominent among his various contributions to linguistics, in that quantitatively-represented variables quite efficiently account for, and possibly predict, actual language use.

4.2. Correspondence between the Generalized Matching Law and the Logistic Regression Model

Logistic regression analysis conceivably contributes to an investigation of the relationship between mere exposure effect and the generalized matching law. Logistic regression is a method often used in medical statistics, biology, and sociolinguistic studies (Matsuda, 1993), and is expressed as follows:

$$\log\{p1 / (1 - p1)\} = Z, \tag{9.1}$$

where *Z* is a linear function, the element *p1* refers to the probability of choosing Alternative 1, and *1–p1* describes the probability of choosing Alternative 2 in two-alternative forced-choice tasks. The ratio of the difference in the probabilities between the two options, i.e. *p1/(1–p1)* in Equation (9.1), is called *odds*. Equation (9.1) can be expanded to Equation (9.2) as follows:

$$p1 = 1 / \{1 + \exp(- Z)\}, \tag{9.2}$$

which yields probability values between 0 and 1, never allowing values less than 0% or equal to / greater than 100%.

When the response frequencies of Alternatives 1 and 2 are replaced with *R1* and *R2*, the sum of response frequencies *N* is expressed as *R1+R2*, as in *N=R1+R2*. Since *p1* refers to the

probability of choosing Alternative 1, *p1* can be defined as *R1/N* and *p2* as *R2/N*. Thus, the ratio of probabilities is expressed by the odds represented by Equation (10).

$$p1 / (1 - p1) = (R1/N) / (R2/N) = R1 / R2, \quad (10)$$

Yokoyama (2006b) shows that Equation (9.1) becomes identical to Equation (8), i.e. the generalized matching law, when *R1/R2* in Equation (10) substitutes *p1/(1–p1)* in Equation (9.1) and *S log (r1/r2)+log B* in Equation (8) replaces *Z* in Equation (9.1) as follows:

$$\log \{p1 / (1 - p1)\} = \log (R1/R2) \quad \text{and} \quad Z = S \log (r1/r2) + \log B,$$
$$\log (R1/R2) = S \log (r1/r2) + \log B, \quad (8)$$

where Equation (8), i.e. the generalized matching law, is a form of logistic regression model. It may be applicable to a wide range of phenomena across various fields of science, gathering evidence from animal behavior, economic, and ecological studies. However, no previous research seems to have pointed out the comparability between logistic regression models and the generalized matching law.

4.3. The Generalized Matching Law and Mathematical Linguistics in Japanese

Yokoyama (2006b) examines the applicability of the generalized matching law in mathematical linguistics, using a preference judgment task, in which the participants choose which of a pair of kanji variants they prefer. The stimuli were pairs of kanji variants, such as “桧” vs. “檜”. The generalized matching law as in Equation (8) is represented by Equation (8a) as follows:

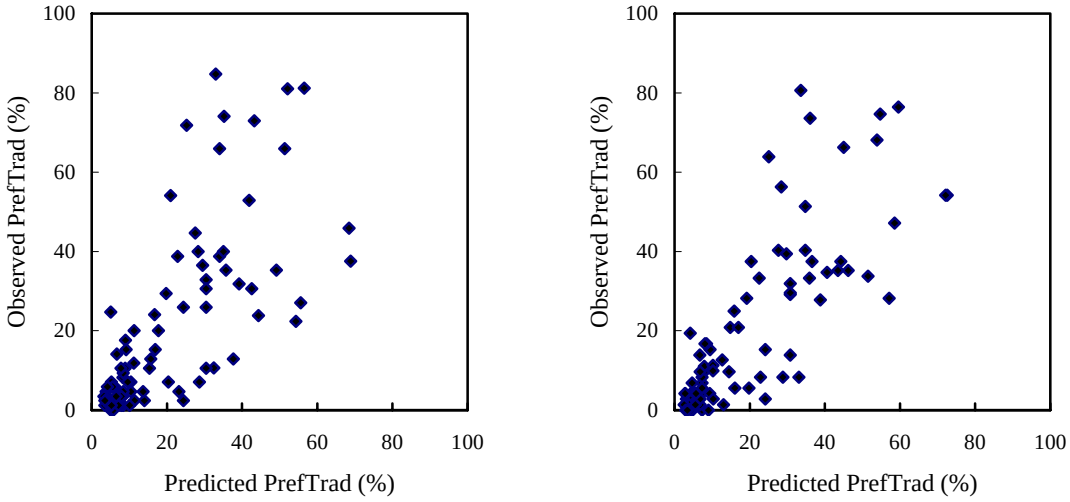
$$\log (R1/R2) = S \log (r1/r2) + \log B = a (FamTrad - FamSimp) + b, \quad (8a)$$

where the numbers of participants who chose the traditional variant and the simplified variant are represented as *R1* and *R2* respectively. Comparability between Equations (8) and (8a) is explained as follows:

$$\begin{aligned} a (FamTrad - FamSimp) + b &= a \{[K \log (FreqTrad) + C] - [K \log (FreqSimp) + C]\} + b \\ &= a K \{\log (FreqTrad) - \log (FreqSimp)\} + b \\ &= a K \log (FreqTrad / FreqSimp) + b \\ &= S \log (r1/r2) + \log B, \end{aligned}$$

where the frequency of the traditional variants, i.e. *FreqTrad*, is defined as *r1*, and frequency of the simplified variants, i.e. *FreqSimp*, as *r2*. The logarithm of the frequency ratio of the characters, i.e. log*(r1/r2)*, is referred to as *exposure relativity*.

Yokoyama (2006b) computes the logarithm of ratio of reinforcer frequencies, i.e. log*(r1/r2)* in Equation (8), based on the frequency data from a newspaper corpus, and estimates the values of the parameter *S* and *log B* by the least square method. It should be noted that the model of Yokoyama (2006b) is an innovative contribution, for studies on mere exposure effect, originally proposed by Zajonc (1968), have not previously presented any specific models. As discussed, the model describes preference behavior in natural language quite reliably, although it still awaits further empirical validation with applicable data.



(a) Tokyo [r^2= .517] (b) Kyoto [r^2 = .646]

Figure 2. Predicted and observed preference for traditional variants

Yokoyama (2006c) estimates the parameters in the Tokyo group by applying the method of maximum likelihood estimation to the model as follows (p<.01, df=1):

$$\log (R1/R2) = 0.294 \log (r1/r2) - 0.301, \quad (11)$$

$$p1 = 1 / \{ 1 + \exp[- 0.294 \log (r1/r2) + 0.301] \}, \quad (12)$$

which expresses the probability of choosing the traditional variant of the pair. Although the maximum likelihood estimation is commonly applied in various fields, its application to a study on variant preference behavior in Japanese kanji recognition is a novel exploration.

The results reveal that the model accounted for 51.7% (r=.719, p<.01, df=84) and 64.6% (r=.804, p<.01, df=84) of the variance of the observed data in the Tokyo and Kyoto groups respectively. This predictive power is considered significantly strong in studies on natural language. Figure 2 shows the correlation between the predicted probabilities and the observed responses of *PrefTrad*.

An inter-group difference of 15% was observed in the accountability between the Tokyo and Kyoto groups (Yokoyama, 2006c). This difference may be attributable to the types and frequencies of kanji characters to which the participants were exposed in their daily lives in the two different locations. Another factor which may have contributed to the inter-group differences is the genders of the participants. The Tokyo group only consisted of females, while the Kyoto group included twenty males (30% of the group). Although these two factors may be responsible for the inter-group differences, further investigation of this point is necessary.

5. Conclusion

The present paper has reviewed a series of related studies and proposed a model which accounts for performance in the preference judgment task in Japanese kanji recognition. The analysis has indicated the strong predictive power of the model of Yokoyama (2006b) with empirical data from the preference judgment task of Yokoyama (2006a) and Yokoyama

(2006b). For example, the model accounts for 64.6% of the variance of the observed responses in the Kyoto group, based on the frequency data from the Asahi Newspaper corpus. The models discussed in the paper are conceptually straightforward, providing opportunities for immediate applications. It should also be noted that they are quite efficient, in that they provide reliable accountability using only frequency data from newspaper text, though newspapers only constitute a small portion of those activities which involve written language.

The basic principle of the mere exposure effect theory mentioned earlier explains this phenomenon quite well. Studies including but not limited to Zajonc (1968) and Kunst-Wilson & Zajonc (1980) assert that repeated exposure to unfamiliar stimuli increases familiarity, and that, as a consequence, preference for such items increases. Provided that frequency data from newspapers represents the actual use of the given orthography, highly frequent items in newspaper corpora are likely to be high-exposure items in actual written communication. Given that repeated exposure to certain items increases preference, high- frequency characters, i.e. highly exposed characters, are more likely to be preferred than less-exposed characters. In addition, preferred items are likely to be used more frequently, further increasing their exposure. In short, the three variables, i.e. actual language use, frequency of exposure, and elevated preference, are mutually dependent, constantly affecting one another. The results of the present study as well as the empirical evidence from Yokoyama (2006a) and Yokoyama (2006b) support such a theoretical framework, describing the role of mere exposure effect and the applicability of the generalized matching law.

The model with strong predictive power included the familiarity difference between the traditional and simplified variants as the explanatory variable. This is probably because the participants in the study compared their degrees of familiarity with the traditional and simplified variants and chose the more familiar items as they performed the task, for the task *per se* was to compare and select one of the two alternatives at a participant-controlled speed. This account leaves room for replication studies for methodological issues to be examined. Another issue is the validity of newspaper corpora as representative of actual language use in general. Further research on this issue must await corpora that represent the full diversity of orthographic characteristics and types of communication in written language.

Acknowledgement: The authors would like express gratitude to Dr. Katsuo Tamaoka for his efforts and encouragement while exchanging views during the various stages of manuscript preparation. We also appreciate the meaningful feedback provided by anonymous reviewers. The paper is based on studies published in *Mathematical Linguistics* Volume 25, Numbers 4 and 5. The authors acknowledge the contributions of the Mathematical Linguistic Society of Japan, which celebrates its 50th anniversary this year. We are especially grateful for the continuous contributions to mathematical linguistics in Japan of Dr. Sizuo Mizutani, the founder of the Mathematical Linguistic Society of Japan and a former member of the National Institute for Japanese Language, who celebrates his 80th birthday this year.

References

Asahi Shinbun-sha (1994). *CD-HIASK 1993 Asahi Shinbun Database*. Tokyo: Kinokuniya & Nichigai Associates.

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*.

Chikamatsu, N., Yokoyama, S., Nozaki, H., Long, E., & Fukuda, S. (2000). A Japanese logographic character frequency list for cognitive science research. *Behavior*

Research Methods, Instruments, and Computers, No. 32, Vol. 3, 482-500.

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

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

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior, 4, 267-272*

Kunst-Wilson, W.R., & Zajonc, R.B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science, 207, 557-558*.

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

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

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

Sasahara, H. & Yokoyama, S. (2000) Familiarity with kanji variants and user preference. *Japanese Linguistics, 8, 110-125*.

Wardhaugh, R. (1986) *An introduction to sociolinguistics*. Oxford: Blackwell Publishers.

Yamaguchi, T., & Ito, M. (2006). An experimental test of the ideal free distribution in humans: The effects of reinforcer magnitude and group size. *The Japanese Journal of Psychology, 76, 547-553.*

Yokoyama, S. (2006a). Can we predict preference for kanji form from newspaper data on character frequency? *Mathematical Linguistics, 25, 181-194.*

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

Yokoyama, S. (2006c, July). Corpus data and prediction of language use by the logistic regression analysis. Paper presented at the research meeting of the National Research for Japanese Institute, Tokyo.

Yokoyama, S., Sasahara, H., Nozaki, H., & Long, E. (1998) *Study on the use of kanji in electronic-media newspapers (Shinbun denshi media no kanji)*. Tokyo: Sanseido.

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