Logistic regression model of preference and familiarity in letter perception

Shoichi Yokoyama

National Institute for Japanese Language

Key words: letter perception, preference, familiarity, mere exposure effect, logistic regression model, Fechner's law, generalized matching law, corpus, maximum likelihood estimation

Summary

In Japanese, pairs of Kanji letters share the same meaning and pronunciation but exhibit varieties in their visual forms, for example "檜 – 桧", they are called *variants*. Variants are commonly found in Japanese, they can often be characterized by a pairs of a "traditional" and a "simplified" forms. For example, a letter representing "cypress" can be transcribed with the traditional form "檜" and the simplified form "桧", both of which are pronounced the same.

The study discussed a kind of regression model to account for and to predict preference from familiarity in Kanji. I proposed that probability of which alternative of Kanji variants is preferred in 2-alternative forced-choice task can be decided with a linear function in mind expressed as follows:

$$Z = a\ (\mathit{FamTrad} - \mathit{FamSimp}) + b, \tag{A}$$

where *FamTrad* stands for familiarity of traditional variants and *FamSimp* for familiarity of the counterpart simplified variants. The study should link this linear function with logistic regression model. It is considered as an efficient way to explain phenomena that can be represented by 2-alternative, such as positive vs. negative. The logistic regression model is expressed as follows:

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

where *Z* is a linear function, the term *p1* refers to the probability to choose Alternative 1, and *1–p1* describes the probability to choose Alternative 2 in forced-choice tasks.

$$\log \{p1 / (1 - p1)\} = a\ (\mathit{FamTrad} - \mathit{FamSimp}) + b, \tag{C}$$

This study estimated the parameters by applying the method of maximum likelihood estimation to equation (C).

Mere exposure effect as familiarity effect

It is well known that previous studies have verified highly positive correlations between familiarity and preference. However, there is not any similar evidence for the relationship between familiarity and preference in letter perception. In social psychology, Zajonc (1968) proposed a "mere exposure effect", showing that repeated exposure to unfamiliar items increases favorability for them in preference judgments. He suggested that high-familiarity words have a tendency to be more preferred than low-familiarity words, observing that these effects are a function of the *logarithm* of exposure frequency.

Preference in *Kanji* character perception

In Japanese, there are some kanji characters share the same meaning and the same pronunciation but which exhibit different manifestations in their visual shapes. For example, the character for *hinoki* 'cypress' has two *variants*, 檜 and 桧. Both variants are common in modern written Japanese, and are often realized by a pair which has a "traditional" and a "simplified" form. For example, the kanji representing *hinoki* 'cypress' can be written with the traditional form 檜 or the simplified form 桧, and both are pronounced the same, that is, as *hinoki*. The difference between the "traditional" and the "simplified" variants only lies in their graphic representations. Both visual forms are real written words which can be observed in naturally occurring text, and are familiar to all Japanese native speakers.

Yokoyama, Sasahara, and Toyama (2006) employed a two-alternative forced-choice task, in which the participants were presented with pairs of kanji variants and asked to choose the item in each pair that was preferable. The participants were instructed to perform the tasks under the assumption that they were word-processing with digital tools such as computers or cell phones. There were approximately 200 participants, and all were college students in the Tokyo area. The data indicated that the participants' preference for variants was not attributable to graphic complexity or to historical reasons, but to the frequencies of the given variants. Yokoyama (2006) and Yokoyama and Wada (2006) further analyzed the contribution of frequency, introducing a psychophysical model, based on Fechner's law, that accounted for their performance in the preference judgment task by referencing frequency data obtained from newspaper corpora.

Preference, familiarity, and logistic regression model

I propose a regression model to account for and to predict preference based on familiarity in Japanese kanji perception. I think that the probability of the preferred alternative in a two-alternative forced-choice task can be explained with a linear function expressed as follows:

$$Z = a\,(FamTrad - FamSimp) + b, \tag{1}$$

where *FamTrad* stands for familiarity of traditional variants and *FamSimp* stands for familiarity of the counterpart simplified variants.

This study links this linear function with a logistic regression model, which is in fact commonly applied to studies in medicine and the biological sciences. It is widely regarded as an efficient way to explain phenomena that can be represented by two alternatives, such as positive vs. negative. The logistic regression model used in variation theory studies is typically expressed as follows:

$$\log\,\{p1 \,/\, (1 - p1)\} = Z\text{ ; therefore} \tag{2}$$

$$p1 = 1 \,/\, \{1 + \exp\,(-\,Z)\}, \tag{2a}$$

where Z is a linear function, the element $p1$ refers to the probability of choosing Alternative 1, and $1\text{–}p1$ describes the probability of choosing Alternative 2 in forced-choice tasks. The ratio of the difference in the probabilities between the two options, i.e., $p1\,/\,(1\text{–}p1)$ is referred to as the *odds*, and the logarithm of the odds, i.e., $log\{p1\,/\,(1 - p1)\}$ is called the *logit.*

$$\log\,\{p1 \,/\, (1 - p1)\} = a\,(FamTrad - FamSimp) + b, \tag{3}$$

It is well known that a logistic curve is very similar to the normal cumulative curve. This study will use the model-based equation (3) above to for reveal the relation between familiarity and preference. The first aim of this study is therefore to examine the hypothesis that the probability of the preferred alternative in a two-alternative forced-choice task can be explained with a simple linear function, whereby familiarity is the mental filter which compares one alternative with the other.

Experiment

I examined the if the probability of the preferred alternatives in a two-alternative forced-choice task can be described by a simple non-linear function, in which familiarity is compared for one alternative to the other, as expressed by Equation (3). This experiment also examined whether regional differences are observed in the preference judgment task with character variants having been conducted in the two distinct regional areas, i.e. Tokyo and Osaka-Kyoto regions, in written languages, because a clear regional variability is known to exist in spoken languages between the two regions.

Participants

The participants consisted of two groups, all of whom were college students. A total of 85 participants in the Tokyo group participated in the experiment in 1996, and 72 in the Kyoto group participated in 1998. Tokyo and Kyoto were selected to represent two diverse geographical regions; the Tokyo group represented the Kanto area around Greater Tokyo and the Kyoto group represented the Kansai area around Greater Osaka. Both are major cities in the eastern and western Japan, respectively.

Materials

The original 263 pairs of variants were selected based on the frequency of their usage in Japanese. 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 table (hence the notation as 83-JIS). The original 263 pairs were narrowed down to 86 pairs of variants, each pair consisting of a traditional kanji character and a simplified character (see Yokoyama and Wada, 2006), and Table 1 presents the actual stimuli used in the task. Presentation order was randomized in the task, as was the location of the stimuli in left or right positions.

Procedure

The purpose of the preference judgment task was to examine which one of the paired variants, i.e., the traditional or the simplified variants, was preferred in variant selection by native speakers of Japanese. The task was a two-alternative forced-choice task, in which the participants were asked to suppose that they were engaged in the task of word-processing; they were instructed to choose the variant, i.e., either traditional or simplified, which they preferred. Word processing was specifically chosen as the context of the task in order to minimize the effects of non-target variables, such as efficiency in hand-writing.

Familiarity data

Familiarity ratings for each character were obtained from character frequency data by Amano, Kondo, and Kakehi (1995) and Kondo and Amano (1999). These familiarity data summarize the means of subjective ratings of familiarity by 24 participants who were asked to rate the subjective familiarity of kanji characters on a 7-point Likert scale, ranging from "completely unfamiliar" to "very familiar". Table 1 shows the familiarity data referenced in the study.

Data on visual complexity

Visual complexity of items was obtained from the data by Kondo and Amano (1999), in which subjective judgment of visual complexity was measured by a 7-point Likert scale. Table 1 shows the visual complexity data referenced in the study.

Stroke number of the characters

Stroke numbers of the items were obtained from a dictionary. Table 1 shows the stroke number data referenced in the study.

Table 1. Examples of differences in familiarity[1], preference for traditional variants, visual complexity differences, and stoke number differences

variant pairs (Trad / Simp)		Familiarity difference	Preference for traditional variants		Visual complexity difference	Stroke number difference
			Tokyo 1996	Kyoto 1998		
壺	壷	0.96	74.1	88.9	-0.09	1
螢	蛍	-0.50	54.1	37.5	0.83	5
鶯	鴬	0.87	65.9	58.3	0.92	5
會	会	-2.45	4.7	2.8	1.50	7
檜	桧	0.21	71.8	70.8	0.66	7
觀	観	-3.29	0.0	2.8	1.16	6
灌	潅	2.17	84.7	90.3	0.30	6
狹	狭	-2.41	17.6	15.3	0.21	1
堯	尭	0.17	31.8	33.3	0.67	4
區	区	-3.92	1.2	0.0	0.92	7
歐	欧	-2.59	24.7	5.6	0.87	7
經	経	-3.00	5.9	1.4	0.96	2
頸	頚	2.00	81.2	83.3	0.08	2
儉	倹	-2.33	7.1	4.3	0.58	5
顏	顔	-3.00	0.0	0.0	0.38	0
國	国	-1.71	10.6	6.9	1.04	3
爾	尓	1.17	37.6	62.5	2.00	9
邇	迩	0.16	27.1	45.8	1.71	10
壽	寿	-2.41	2.4	1.4	1.54	7
濤	涛	0.54	38.8	40.3	1.05	7
檮	梼	-0.33	25.9	33.8	1.21	7
孃	嬢	-1.50	12.9	16.7	0.59	4
飮	飲	-3.46	1.2	2.8	0.38	1
眞	真	-2.16	15.3	4.2	0.91	0
愼	慎	-2.24	15.3	12.5	0.71	0
槇	槙	-0.37	44.7	40.3	0.71	0
盡	尽	-2.92	2.4	2.8	1.46	8
儘	侭	-0.12	30.6	37.1	1.17	8
數	数	-3.50	1.2	0.0	1.12	2
藪	薮	-0.04	40.0	38.9	0.71	2
錢	銭	-2.25	9.4	8.3	0.96	2
賤	賎	1.37	65.9	72.2	0.63	2
曾	曽	-1.50	20.0	27.8	0.63	1
澤	沢	-0.66	38.8	30.6	1.00	9
驛	駅	-3.63	10.6	5.6	1.25	9
諫	諌	0.54	40.0	51.4	0.50	1

[1] Differences were calculated by subtracting the values for simplified variants from those for the traditional counterparts.
eg. Familiarity differences = *Famtrad – Famsimp*

Results

Regional differences were not systematically observed in the preference ratio for the variant pairs. Logistic regression analysis was performed with the percentages of preference for the variants, as explained in equation (3). Table 2 shows the values of parameters, estimated by the method of maximum likelihood estimation (Collett, 2003). Fittings for parameter values was tested by the Wald test, and all were significant ($p < .01$, $df = 1$). The predicted probabilities of the preference for the traditional variants were computed as follows:

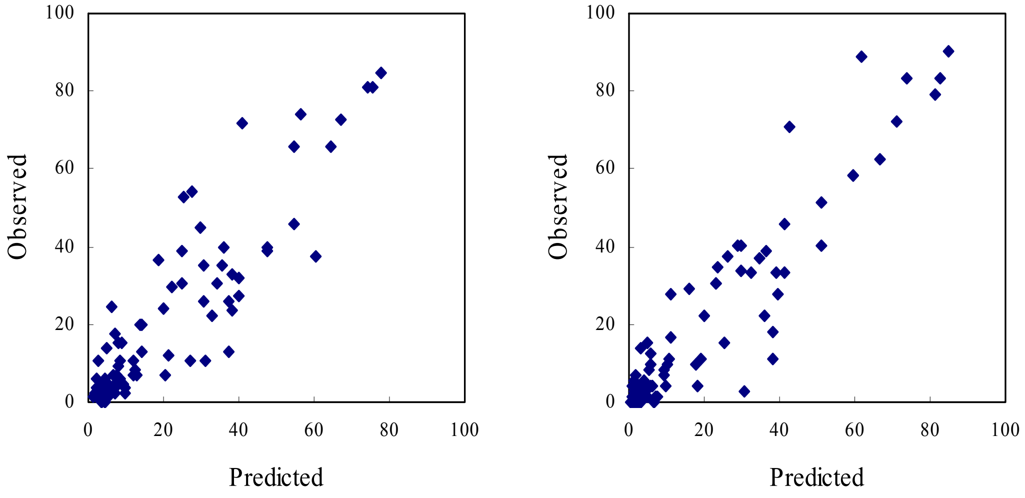
$$p1 = 1 \,/\, \{1 + \exp\left[\, - a\,(FamTrad - FamSimp) - b \,\right]\}, \qquad (3a)$$

where $p1$ stands for the predicted probability of the preference for the traditional variants. It might also be noted that regional differences were very small between Tokyo and Kyoto in the estimated parameters in Table 2.

Table 2. Correlations between predicted and observed preference, and estimated parameter values in Equation (3) based on familiarity data

Region area	*r*	*a*	*b*
Tokyo 1996	.900	0.835	-0.547
Kyoto 1998	.935	1.033	-0.517

The predicted and observed preferences for traditional variants were compared. Table 2 shows the values of significant and strong correlations between the observed and predicted preference in both Tokyo and Kyoto, yielding Pearson r ranging from .900 to .939 ($p < .01$, $df = 84$). Such a strong correlation plausibly indicates both preference and familiarity are psychological variables related to each other.



(a) Tokyo 1996

(b) Kyoto 1998

Figure 1. Observed and predicted preference percentage for traditional variants based on familiarity.

Figure 1 shows the scatter-gram of the correlation between the predicted and observed preference for traditional variants. Table 2 also shows that regional differences were very small between Tokyo and Kyoto in *r* values.

Discussion

This study examined the extent of correlations between familiarity and preference in kanji character perception by using a logistic regression model. The results showed that the model reliably accounted for the performance in two major metropolitan areas in Japan, Tokyo and Kyoto. The predictive power of the logistic regression model was extremely strong, with the fitting of preference judgments between observed data and predicted data from familiarity ratings showing all *r* correlation coefficients over .900. The basic principle of the mere exposure effect theory mentioned in the opening sections of this paper appears to explain this phenomenon quite satisfactorily. That is, Zajonc's (1968) suggestion that repeated exposure to unfamiliar stimuli increases familiarity, and that, as a consequence, preference for such items increases, seems borne out by the findings reported here.

Fechner's law bridging frequency and familiarity

The findings of previous studies obviously indicate that exposure frequency increases familiarity, and that familiarity and preference are closely related to each other. Yokoyama and Wada (2006) assumed that exposure to kanji characters, operationally defined as frequency, contributes to preference, defined as performance in the preference judgment task. An explanatory model of familiarity is also available from Fechner's law, as follows:

$$Y = K \log (I) + C, \qquad (4)$$

where Y stands for strength of sensitivity, I for strength of reinforcement, K for the slope, and C for the constant. Based on equation (4), the familiarity of a variant can be expressed as follows:

$$\mathit{Familiarity\ of\ variant} = K \log (\mathit{frequency\ of\ the\ variant}) + C\ , \qquad (5.1.)$$

where familiarity is expressed by a linear function with frequency and *log* refers to natural logarithms with base *e*. By the same token, the familiarity of the traditional and the simplified variants is expressed as follows:

$$\mathit{FamTrad} = K \log (r1) + C\ ;\text{ and} \qquad (5.2.)$$

$$\mathit{FamSimp} = K \log (r2) + C, \qquad (5.3.)$$

where *r1* and *r2* stand for the frequencies of the traditional and simplified variants, respectively. Equations (5.2.) and (5.3.) can substitute *FamTrad* and *FamSimp* in equation (3) as follows:

$$\log \{p1 / (1 - p1)\} = a\ \{\ [K \log (r1) + C] - [K \log (r2) + C]\ \} + b\ ;\text{ therefore}$$

$$\log \{p1 / (1 - p1)\} = a\,K \log (r1 / r2) + b = S \log (r1 / r2) + \log B, \qquad (6)$$

where parameter S refers to the slope of the line representing sensitivity of reaction, and *log B* refers to the intercept representing response bias.

When the response frequencies of Alternatives 1 and 2, i.e., *p1* and *1-p1*, are replaced with *R1* and *R2*, the sum of response frequencies *N* is represented by the odds as follows (Yokoyama, 2006; Yokoyama and Wada, 2006):

$$p1 / (1 - p1) = (R1 / N) / (R2 / N) = R1 / R2, \qquad (7)$$

where *p1* refers to the probability of choosing Alternative 1. Therefore, *p1* can be defined as *R1/N* and *p2* as *R2/N*. Then equation (6) can be written as follows:

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

Generalized matching laws in animal behavior

Baum (1974) showed that the basic principle of the generalized matching law expresses the relationship between the reinforcement and response allocation in a model that is very similar to the logistic equation seen in (8) above.

The ratio of reinforcement, i.e., *r1/r2* in equation (8), and that of response allocation, i.e., *R1/R2*, may be exemplified by a hypothetical experiment in the following way. Suppose that pigeons or rats in cages are rewarded with food by pushing levers called Levers 1 and 2. . 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 to as the ratio of reinforcement, i.e., *r1/r2*. The frequencies of lever-pushing behavior by the animals are represented by *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 the ratio of response allocation is concisely expressed by the ratio of reinforcement expressed in equation (8) above.

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). The model of the ideal free distribution theory is identical to the equations in (6) and (8) above, 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 provides satisfying explanations for empirical evidence found across different fields of study.

Application to corpus linguistics

In corpus linguistics and mathematical linguistics, the ratio of reinforcement is computed by using the frequency data of variants based on corpora. In terms of this study, variants were pairs of traditional and simplified forms, and thus the frequency of the traditional variant found in newspaper corpora was defined as *r1* and that of the counterpart simplified variant was defined as *r2*. The ratio of responses was obtained from responses in a preference judgment task in which the participants chose one of the paired variants, i.e,either the traditional form or the simplified form. The number of participants who chose the traditional variant in the preference judgment task was represented by *R1,* while the number of participants who chose the counterpart simplified variant was represented by *R2*. Thus equation (8) can be applied as follows:

$$\log\,(R1\,/\,R2) = S\log\,(\,FreqTrad\,/\,FreqSimp\,) + \log B, \qquad (8a)$$

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*.

Yokoyama (2006) and Yokoyama and Wada (2006) demonstrated that kanji frequency data from the Asahi newspaper corpus explained the performance in a preference judgment task by applying the generalized matching law. Yokoyama and Long (2007) showed that when the predicted and the observed response ratios were compared, significant correlations with *r*= .65 or greater emerged, based on the frequency data obtained from corpora that included newspaper, encyclopedia, and literary texts. This predictive power is considerably strong when applied to studies of natural language.

Circular model

The role and predictive power of frequency in language use can be explained by a circular model as illustrated in Figure 2. Language policies and social frequency represent the social phenomena observable in the given community. The two-way arrow represents mutual contribution between language policy and social frequency. Social frequency contributes to exposure frequency in a one-way manner, as represented by the one-way arrow. Exposure frequency is the degree of exposure, at which individual language users are exposed to the given language forms in the given community. Exposure frequency mediates the contribution of the social frequency to individuals' mental lexicon. In other words, it is the gateway for social phenomena to get integrated into the individuals' psychology. Individual's psychology is typically represented by mental lexicon in the figure, which includes familiarity, preference, and utility. Positive behavior of mental lexicon, increased familiarity for example, leads to increased reproduction of the given language forms in the series of chain reaction as illustrated in Figure 2, consequently promoting the social frequency even more. Such a circular relationship provides social frequency with increasingly stronger impacts in the chained and circular phenomena. The increasingly stronger impacts of social frequency allow a heavier and increasingly more reliable role of frequency data, which is amplified in a non-linear fashion due to the circular flow of the phenomena involving multiple variables.

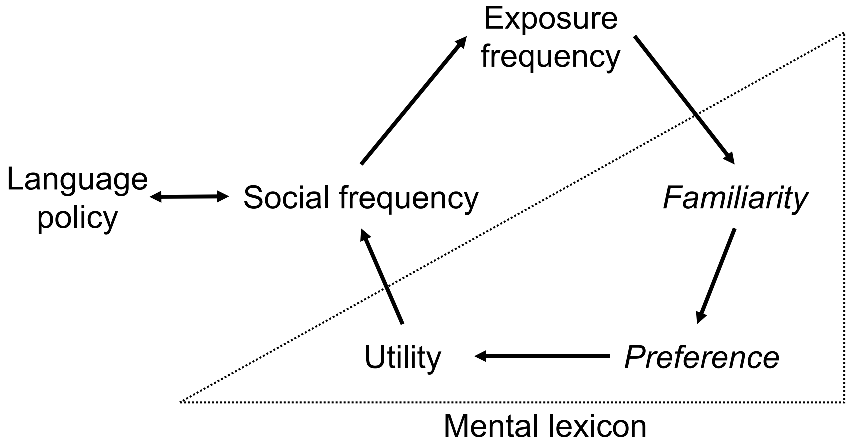


Figure 2. Circular model by Yokoyama (2007)

Acknowledgement

The author thanks Professor Kess at University of Victoria, Canada, for his valuable comments on earlier versions of the manuscript. He also acknowledges the editorial assistance of Yukiko Wada in the preparation of English manuscript.

References

Amano S., Kondo T., & Kakehi K. (1995). Modality dependency of familiarity ratings of Japanese words. *Perception & Psychophysics*, **57**, 598-603.

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.

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

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.

Kondo, T., & Amano, S. (1999). *Lexical properties of Japanese*, **5**, Character, NTT Database Series.

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.

Yokoyama, S., & Long, E. (2007). Logistic Regression Analysis of Preference for Kanji variants: Maximum Likelihood Estimation Based on Text Corpora. *Mathematical Linguistics*, **26**, 19-30.

Yokoyama, S., Sasahara, H., & Touyama, H. (2006). Preference and familiarity of kanji variants in Japanese reading and writing communication：Measurement of reliability with retest method. *The Japanese Journal of Language in Society*, **9**, 16-26.

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