Estimation of Mental Lexicon Size with Word Familiarity Database

Shigeaki Amano & Tadahisa Kondo

NTT Basic Research Laboratories

3-1 Morinosato Wakamiya, Atsugi, Kanagawa 2430198, Japan

ABSTRACT

A familiarity database was developed for about 80,000 Japanese words of which familiarity scores were rated by 32 Japanese adults using a 7-point scale in auditory, visual, and audio-visual modalities. Auditory, visual, and audio-visual stimulus words were selected from the database according to their word familiarity for size estimation of the mental lexicon. Sixty Japanese adults participated in a two-alternative forced-choice task (Know–Don't know) for the stimulus words. The size of the mental lexicon was estimated as the number of words of which familiarity is above a particular word corresponding to 50% point on the fitted logistic curve to "know"-response probability of the stimulus words. The estimated size was about 68,000 for auditory words, and about 66,000 both for visual and audio-visual words when homophones and homographs were included. The results suggest that very small difference in the mental lexicon size among modalities.

1. INTRODUCTION

The size of the mental lexicon has been estimated by several methods. With dictionary sampling method, for example, the size is estimated by multiplying the number of words in the population (i.e., dictionary or database) and "know"-response probability for randomly-sampled words from the population (e.g., Gillette [1]; Hartmann [2]; Seashore & Eckerson [3]). With familiarity or frequency sampling method, the population is divided into different familiarity or frequency ranges and the size is estimated by summing up the number of words multiplied by "know"-response probability for randomly-sampled words across the ranges (e.g., D'Anna, Zechmeister, & Hall [4]). With exhaustive counting method, the size is estimated by counting the words of which familiarity value is greater than a particular value (e.g., Morioka [5]; Nusbaum, Pisoni, & Davis [6]).

However, there is a large discrepancy between the estimates of the size of the mental lexicon, as shown in Table 1. The discrepancy is thought to be partly due to methodological differences, and partly due to differences in the population size. This is because the population size provides the upper limit of the estimated size, and therefore, the larger the population size is, the larger the size of a mental lexicon becomes. Another problem is that the estimations have been made only on the basis of visual words (i.e., written words) and not on the basis of auditory words (i.e., spoken words) or audio-visual words (i.e., audio and visual words are simultaneously presented). The size might be different between different modalities.

In the present study, a word familiarity database in Japanese was firstly developed for a large number of words in three modalities: audio, visual, and audio-visual. This database is a new version of the previous word familiarity database (Amano, Kondo, & Kakehi [7]). Secondly, using the new database, the mental lexicon size of Japanese young adults was estimated in audio, visual, and audio-visual modalities using a new familiarity-sampling method which is based on logistic-curve fitting to "know"-response probability of sampled words. Possible size differences were also investigated between different modalities. Homophones and homographs were taken into consideration in the estimation, because they are abundant in Japanese.

Study	Estimated Size of Mental Lexicon	Method	
		Sampling	Variable
Nusbaum et al. [6]	14,418(19,750)	Exhaustive	Familiarity
D'Anna et al. [4]	16,785(26,901)	Random	Familiarity
Morioka [5]	30,664(37,970)	Exhaustive	Familiarity
Gillette [1]	127,800(450,000)	Random	-
Seashore & Eckerson [3]	155,736(454,088)	Random	-
Hartmann [2]	238,620(454,088)	Random	-

Table 1: Estimated sizes of the mental lexicon and estimation methods used in the previous studies. The size of population used for each study is represented in parentheses.

2. EXPERIMENT 1

Experiment 1 was conducted to develop a word familiarity database for a large number of Japanese words. Figure 1 shows a flow chart for Experiment 1. Pretest 1 and 2 were conducted to exclude inappropriate accent types and orthographies.

2.1. Pretest 1

Objective. The objective of Pretest 1 is to obtain rating scores of accent appropriateness for Japanese words.

Subjects. Ten Japanese adults (6 males and 4 females) in their twenties with Tokyo-dialect participated in the experiment. They and their parents were born and grew up in Tokyo area.

Stimuli. About 80,000 words in a Japanese dictionary [8] were used. Stimuli consisted of lists of words with their orthography, their pronunciation written in *Katakana*, and their accent types.

Procedure. Subjects were trained to identify accent type before this experiment. The training took two days until all

subjects can correctly identify the accent type. A half of the subjects received the stimulus lists in normal order and the other half of the subject received the lists in reversed order to reduce a context effect. The subjects rated accent appropriateness of each accent type for each word in the list using a 5-point rating scale (1: inappropriate – 5: appropriate). If the subject found neither of accent types were appropriate for the word, he/she was asked to added an alternative accent type and rated its appropriateness using the scale.

Results. Mean rating scores of accent appropriateness were obtained for each accent type of each word. The number of subjects who rated the each accent type was also obtained for each word.

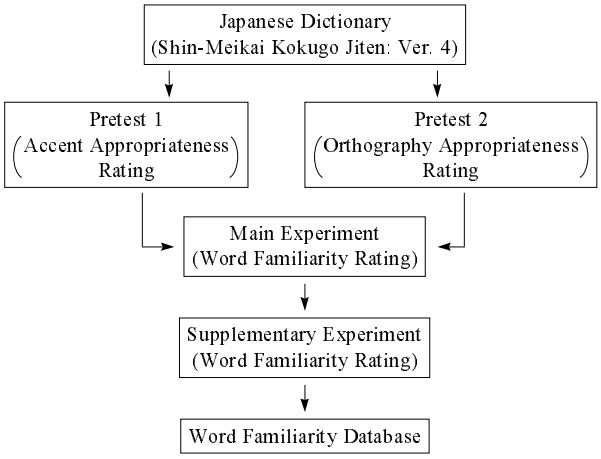

Figure 1: A flow chart for Experiment 1.

2.2. Pretest 2

Objective. The objective of Pretest 2 is to obtain rating scores of orthography appropriateness for Japanese words.

Subjects. Twelve Japanese adults (6 males and 6 females) in their twenties participated in Pretest 2.

Stimuli. About 80,000 words in a Japanese dictionary [8] were used. Each stimulus word was written in the following three types of character strings. 1) a mixture of Kanji, Hiragana, Katakana, and non-Japanese characters such as alphabet (a particular word may have multiple character strings in this type), 2) Hiragana-only, and 3) Katakana-only.

Procedure. Three different types of character strings for each word with a 5-point rating scale (1: inappropriate – 5: appropriate) were presented on a computer screen in a random order. The subjects rated orthography appropriateness of the character strings by choosing one of the numbers on the rating scale using a mouse device. If the subject found neither of the character strings were appropriate for the word, he/she was asked to write down an alternative character string. The alternative character strings were collected across the subjects and they were rated at the end of the experiment by all subjects.

Results. Mean rating scores of orthography appropriateness were obtained for three different character strings of each word.

2.3. Main Experiment

Objective. Objective of this experiment was to obtain word familiarity rating scores in auditory, visual, and audio-visual modalities.

Subjects. Forty Japanese adults (20 males and 20 females) in their twenties participated in this experiment.

Stimuli. The words with appropriate accent types were selected as an auditory word set based on the result of Pretest 1. The selected words had values of more than 30 (maximum: 50) of a product of the word's accent-appropriateness score by the number of rating subjects. Homophones were reduced into one auditory word in the set. There were 62,558 auditory words. The words with appropriate orthographies were selected as a visual word set based on the result of Pretest 2. The selected words had a score of more than 3.75 on a 5-point scale for orthography appropriateness. Homographs were reduced into one visual word in the set. There were 77,036 visual words. These auditory and visual words were paired and used for an audio-visual word set. This pairing provided 89,215 audio-visual words.

Procedure. The stimuli were presented in three different modalities as follows: auditory, visual, and audio-visual presentations. In the auditory presentation, the words were diotically presented to the subject through headphones in 75 dBSPL. In the visual presentation, the words were presented with a 32 x 32 dot font on a computer screen. In the audio-visual presentation, the auditory words and the visual words were simultaneously presented. The presentation order of the stimulus words were randomized for each subject in all types of presentation. At each trial, the subjects rated word familiarity to the stimulus words by choosing one of the numbers on a 7-point rating scale (1: unfamiliar – 7: familiar) appeared on the computer screen using a mouse device. Randomly-selected 9,000 stimulus words were audio-visually presented to the subjects in the training session. After the training session, the subjects rated word familiarity in auditory, visual, and audio-visual presentation. The order of the presentations was counter balanced. A posttest was conducted in all types of presentation to check rating stability of the subjects. 3,000 words were randomly selected for the each presentation of the posttest. The procedure of the posttest was the same as the main experiment.

Results. Thirty-two subjects passed the posttest, because they had more than .5 correlation of rating scores between the main experiment and the posttest in all presentation modalities. By averaging their rating scores, word familiarities were obtained for the auditory, visual, and audio-visual modalities.

2.4. Supplementary Experiment

Objective. Some stimulus words in the main experiment were found to be misprepared in their character string and/or their

pronunciation. Supplementary experiment was conducted to obtain word familiarity scores for the misprepared words.

Subjects. Forty Japanese adults (20 males and 20 females) in their twenties participated in the experiment.

Stimuli. Misprepared words in the main experiment were used. There were 770 auditory words, 1,073 visual words, and 1,092 audio-visual words. Filler words were randomly selected from audio, visual, and audio-visual word set to make each word set contain 4,500 words in total. 500 words were selected from fillers for posttest.

Procedure. The procedure was the same as in the main experiment.

Results. Thirty-five subjects passed the posttest, because their correlation coefficients were more than .5 between the rating scores in the supplementary experiment and those in the posttest in all presentation modalities. Using their data, word familiarity scores were obtained for the auditory, visual, and audio-visual modalities.

2.5. Word Familiarity Database

The main and supplementary experiments provided word familiarities for 62,556 auditory words, 77,040 visual words, and 89,224 audio-visual words. These familiarities were registered in a word familiarity database with their orthography, pronunciation, and accent type. Table 2 shows the number of words which are categorized as whether a word has homophones and/or homographs. In this study, the homophones are defined as words with the same pronunciation with different orthography in the familiarity database even if they represent the same meaning. The homographs are defined as words with the same orthography with different pronunciation (including different accent type) in the familiarity database even if they represent the same meaning. Table 3 shows the accumulative number of words as a function of word familiarity. Because Table 2 shows that there are many homophones and homographs, the counting was conducted both for including all homophones and homographs and for reducing homophones and/or homographs into one word. This reduction is not available for audio-visual words because they do not have any homophones and homographs by definition.

		Visual Word		Total
		Unique	Homo-graphic	
Auditory Word	Unique	33,972	11,419	45,391
	Homo-phonic	32,266	11,567	43,833 (17,175)
Total		66,238	22986 [10,802]	89,224 (62,566) [77,040]

Table 2: The number of words in the word familiarity database. The number of auditory words with reducing homophones into one word is represented in parentheses. The number of visual words with reducing homographs into one word is represented in square brackets.

Lower Limit of Word Familiarity	Auditory Word		Visual Word		Audio-Visual Word
	With All Homo-phones	Reducing Homo-phones into one word	With All Homo-graphs	Reducing Homo-graphs into one word	
6.0	4,540	3,147	5,240	4,501	4,562
5.0	35,755	23,324	30,520	25,750	28,764
4.0	55,957	36,653	51,374	43,322	49,383
3.0	70,684	47,210	68,805	58,537	67,018
2.0	86,307	59,873	83,866	72,131	83,680
1.0	89,224	62,566	89,224	77,040	89,224

Table 3: The accumulative number of words in the word familiarity database as a function of the lower limit of word familiarity. The auditory, visual, and audio-visual familiarity was respectively used for counting auditory, visual, and audio-visual words.

3. EXPERIMENT 2

Objective. The objective of this experiment is to estimate the mental lexicon size in auditory, visual, and audio-visual modalities.

Subjects. Sixty Japanese adults (30 males and 30 females) participated in the experiment. They were 19 to 29 years old (Mean=21.8, S.D.=2.3).

Stimuli. Auditory, visual, and audio-visual word sets were selected from the word familiarity database. Each set consisted of 400 Japanese words which do not have homophones and homographs. Auditory, visual, and audio-visual word familiarities ranged from very low to very high with almost constant intervals in every set. Mean word familiarities were 4.12 (S.D.=1.37), 4.19 (S.D.=1.38), and 4.16 (S.D.= 1.40) for audio, visual, and audio-visual word sets respectively. No significant differences of familiarity were found between any pair of the word sets.

Procedure. The stimuli were presented in three different modalities as follows: auditory, visual, and audio-visual presentations. In the auditory presentation, the words were diotically presented to the subject through headphones in 75 dBSPL. In the visual presentation, the words were presented with a 32 x 32 dot font on a computer screen. In the audio-visual presentation, the auditory words and the visual words were simultaneously presented. The presentation order of the stimulus words were randomized for each subject in all types of presentation modalities. The subject participated in a forced-choice task with two alternatives (Know–Don't know) for the stimulus word by choosing a “Know” or “Don't know” bottom on a computer screen by using a mouse device. The order of the presentations was counter balanced across the subjects.

Results. The ``know''-response probability of each stimulus word was calculated by dividing the number of subjects with ``know''-responses to the stimulus word by the total number of subjects. For each word, the number of the more familiar words than itself were counted in the word familiarity database

either including homophones and homographs or reducing them into one auditory or visual word. By the Interactively Reweighted Least Square (IRLS) algorithm, the logistic curve was fitted to the ``know"-response probability against the number of the more familiar words. A fitting example is shown in Figure 2. The size of the mental lexicon was estimated by calculating the number of the words which corresponds to 50% probability on the logistic curve. The estimated size of the mental lexicon is shown in Table 4.

Modality	Estimating Method	
	Including All Homophones and Homographs	**Reducing Homophones or Homographs into One Word**
Auditory	68,021 (89,224)	45,900 (62,556)
Visual	65,900 (89,224)	56,523 (77,040)
Audio-Visual	66,381 (89,224)	-

Table 4: Estimated size of the mental lexicon. A number in parentheses represents the population size of words used for the estimation.

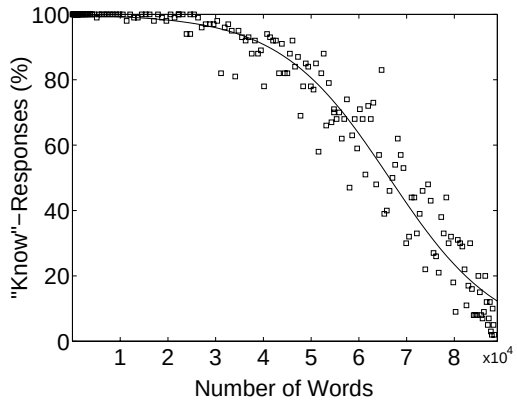

Figure 2: "Know"-response probability for audio-visual words as a function of the number of more familiar words than the stimulus word. Fitted logistic curve is also plotted.

4. DISCUSSION

The new word familiarity database developed here is superior to the previous word familiarity database (Amano, Kondo, & Kakehi [7]) in several features: the number of words (about 89,000 vs. about 62,000), the number of modalities (three vs. two), the number of subjects (32 vs. 11), and the exclusion of inappropriate accent types and orthographies based on pretests. This new database provides practicable estimations of the size of the mental lexicon in three different modalities. The size estimation experiment was conducted with stimulus words sampled from the database. With the familiarity-sampling method with logistic curve fitting, the estimated sizes of the mental lexicon in the audio, visual, and audio-visual modalities were about 66,000 – 68,000 when all homophones and homographs are included in the estimation. The results suggest that a very small difference in the mental lexicon size between modalities. However, when the homophones or homographs are reduced into one word, a large difference was observed. The lexicon size for auditory words (about 45,900) was much smaller than that for visual words (about 56,500). This indicates that homophones and homographs are critical factors for estimating the size of a mental lexicon at least for Japanese in auditory and visual modalities. The exhaustive counting method is applicable to Table 3. It provides almost the same estimated sizes as the familiarity-sampling method with logistic curve fitting, *if the lower limit of word familiarity is set at 3* in Table 3. However, we cannot beforehand know which lower limit should be used. Therefore, the familiarity-sampling method with logistic curve fitting is better than the exhaustive counting method. The estimated size obtained in the present study is much larger than that found by other familiarity-based studies (e.g., D'Anna, Zechmeister, & Hall [4]; Morioka [5]; Nusbaum, Pisoni, & Davis [6]). It is probably because the population size of this study is much larger than that in the other studies. The current estimation is much smaller than the estimates based on dictionary-sampling method (e.g., Gillette [1]; Hartmann [2]; Seashore & Eckerson [3]). It is probably because these studies include derivatives for the estimation. If the derivatives are excluded, the estimated size would be much smaller and similar to the estimates of this study.

5. REFERENCES

1. Gillette, J. M. "Extent of personal vocabularies and cultural control," *The Scientific Monthly* 29: 451–457, 1927.
2. Hartmann. G. W. "A critique of the common method of estimating vocabulary size, together with some data on the absolute word knowledge of educated adults," *The Journal of Educational Psychology* 32: 351–358, 1941.
3. Seashore, R. H., and Eckerson, L. D. "The measurement of individual differences in general English vocabularies," *The Journal of Educational Psychology* 31: 14–38, 1940.
4. D'Anna, C. A., Zechmeister, E. B., and Hall, J. W. "Toward a meaningful definition of vocabulary size," *Journal of Reading Behavior* 23: 109–122, 1991.
5. Morioka, K. "Vocabulary size estimation for graduates from a junior high school," *Annual Report of The National Language Research Institute* 2: 95–107, 1951. (In Japanese).
6. Nusbaum, H. C., Pisoni, D. B., and Davis, C. K. "Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words," *Research on Speech Perception: Progress Report* 10: 357–372, 1984. (Bloomington, IN: Indiana University).
7. Amano, S., Kondo, T., and Kakehi, K. "Modality dependency of familiarity ratings of Japanese words," *Perception & Psychophysics* 57: 598–603, 1995.
8. Kindaichi, K., Shibata, T., Yamada, A., and Yamada, T., *Shin-Meikai Kokugo Jiten (ver. 4)*, Sanseido, Tokyo, 1989.