Word Familiarity Rate Estimation Using a Bayesian Linear Mixed Model

Word Familiarity Rate Estimation Using a Bayesian Linear Mixed Model

Masayuki Asahara
National Institute for Japanese Language and Linguistics, Japan
masayu-a at ninjal dot ac dot jp

Abstract

This paper presents research on word familiarity rate estimation using the 'Word List by Semantic Principles'. We collected rating information on 96,557 words in the 'Word List by Semantic Principles' via Yahoo! crowdsourcing. We asked 3,392 subject participants to use their introspection to rate the familiarity of words based on the five perspectives of 'KNOW', 'WRITE', 'READ', 'SPEAK', and 'LISTEN', and each word was rated by at least 16 subject participants. We used Bayesian linear mixed models to estimate the word familiarity rates. We also explored the ratings with the semantic labels used in the 'Word List by Semantic Principles'.

1 Introduction

Compiling a lexicon is difficult work. In the lexicography field, there are two main types of methodology that are utilized to compile lexicons. One is a corpus-based methodology, which supports the objectivity of the language resources and results. This methodology requires large-scale, balanced corpora to function, which do exist in several languages; for instance, there are several corpus databases for the Japanese language, such as the 'Balanced Corpus of Contemporary Written Japanese' (Maekawa et al., 2014), the 'Corpus of Spontaneous Japanese' (Maekawa et al., 2000) and the 'NINJAL Web Japanese Corpus' (Asahara et al., 2014). In contrast to the corpus-based lexicography, the intuition-based method is more rooted in the subjective perspective of the lexicographer. Nowadays, however, we can perform large-scale experiments that gather enough crowdsourced subjective perspectives to constitute objective linguistic data on individual words.

Generally, a lexicon covers several layers of linguistic features, such as pronunciation, morphological information, part-of-speech or word class, relevant syntactic phenomena, and semantic categories. In addition, the terms in a lexicon include additional features that are used in daily life. One such language resource in Japanese is the 'Word Familiarity Rate', which measures how familiar people are with a specific word by NTT[1](Amano and Kondo, 1999). However, this 'Word Familiarity Rate' experiment was completed more than twenty years ago, and it is therefore possible that the usage and register of words have changed in the intervening years.

In this study, we construct a word familiarity rate database using entries extracted from the 'Word List by Semantic Principles' (『分類語彙表』 Bunrui goihyo, hereafter WLSP) (Kokuritsu_Kokugo_Kenkyusho, 2004). We utilized crowdsourcing to perform a large-scale subjective experiment on 96,557 WLSP entries. We asked the subject participants to rate the familiarity of words along five perspectives: KNOW, WRITE, READ, SPEAK, and LISTEN. The quality of results gathered by crowdsourcing may be lower than that of results collected in a controlled experiment; however, the cost of constructing a crowdsourced study is lower than the cost of conducting an experiment. We utilized a Bayesian linear mixed model (Sorensen et al., 2016) to alleviate noise in the data.

Our work makes the following contributions to the literature:

- We compiled a word familiarity rate database for thesaurus entries.
- We used crowdsourcing via human subject participants to explore word ratings.
- We introduced a Bayesian linear mixed model to this type of rate modelling.

[1]Nippon Telegraph and Telephone Corporation.

Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP, pages 6–14

https://doi.org/10.18653/v1/P17

Table 1: Example Entry from the 'Word List by Semantic Principles'

「昨年」 'Last Year': 1.1642			
Syntactic Category	Semantic Category		
	Top Level	Second Level	Finest Level
体	関係	時間	過去
Nominal Word	Relation	Time	Past Time
1.	.1	.16	.1642

- The word list was taken from the surface forms of WLSP. This enabled us to connect word familiarity rates with the semantic categories in a thesaurus. Kondo et al. (2018) produced a correspondence table between WLSP and UniDic (a lexicon with morphological information). The morphological analyser MeCab enabled us to automatically annotate the familiarity rates using these resources.

- The preceding work introduced the contrast between character-based (WRITE, READ) and voice-based (SPEAK, LISTEN) perspectives. We contributed to the literature by also introducing a new contrast between production (WRITE, SPEAK) and reception (READ, LISTEN) perspectives.

The remainder of this paper is organised as follows. Section 2 presents related work on the 'Word List by Semantic Principles' and the 'Word Familiarity Rate' in Japanese. Section 3 displays the methodology that we used to develop the word familiarity ratings, namely, crowdsourcing and a Bayesian linear mixed model. Section 4 evaluates the results, and Section 5 presents a conclusion and discusses future research.

2 Related Work

2.1 'Word List by Semantic Principles'

The 'Word List by Semantic Principles' (分類語彙表, WLSP) is one of the major thesauri for contemporary Japanese. The first version of the WLSP was released in 1964 by Kokuritsu Kokugo Kenkyusho (Kokuritsu_Kokugo_Kenkyusho, 1964), and a newer, expanded version was published in 2004 (Kokuritsu_Kokugo_Kenkyusho, 2004). Its comma separated value (CSV) file of the expanded version can be used for research purposes.[2]

[2] 200,000 yen (+ tax) for commercial use.

The data include more than 90,000 words with four syntactic categories (nominal word, verbal word, modifer word, and other) and several hierarchical semantic levels. The categories are indicated with a one integer digit to the left of a radix point and with four fractional digits to the right of the radix point. Table 1 shows an example of the word '昨年 (Last Year)', which is assigned a value of 1.1642. Here, the first '1' presents the syntactic part, which is referred to as the 'Nominal Word', while '1642' presents the hierarchical semantic part, as follows: the first digit, '.1', refers to the top-level semantic category 'Relation'; the two digits '.16' refer to the second-level semantic category 'Time'; and the four digits '.1642' refer to the finest-grained semantic category 'Past Time'. These five digits are therefore referred to as the 'WLSP number'. The syntactic categories are 1. Nominal Word, 2. Verbal Word, 3. Modifier Word, and 4. Other (e.g. Conjunction, Interjection, Greeting).

We used all the words as the target words to be annotated for familiarity rates.

2.2 Word Familiarity Rate in Japanese

Preceding work used two methods to estimate the word familiarity ratings: a word frequency-based (objective) and a cognitive experiment-based (subjective) method. The *Nihongo-no goitokusei database* (Amano and Kondo, 1999) includes both objective and subjective data for word familiarity ratings. The data were constructed from 14 years of *Asahi Shinbun* newspaper articles, from 1985 to 1998. They used a morphological analyser, Sumomo, to analyse the articles and split the sentences into words.

The subjective data are cognitive experiment-based. The 40 participants rated word familiarity of three types of stimuli: character-based, voice-based, and both. The participants were chosen based on 'Hyakurakan' (百羅漢), – a Japanese proficiency test – to control their linguistic compe-

tence. The rating score is an integer from 1 (lowest) to 7 (highest), and the number of target entries is 88,569 of character and voice-based stimuli, from 69,084 words. The data gathering was held from September 1995 to July 1996 in the NTT institute. Even though the rating environment was controlled, the estimation of the word familiarity was based on the average of ratings by participants. More sophisticated statistical analysis should be utilised for reducing the subject participant biases.

3 Methodology

3.1 Design

In this section, we present our methodology for constructing a word familiarity rate lexicon at low cost. The word list constitutes 96,557 words taken from the WLSP. We did not prepare any voice data (oral pronunciations) for the lexical entries, but we did cover speech and hearing as two of the following five perspectives:

KNOW: how much do you know about the target word?

WRITE: how often do you write the word?

READ: how often do you read the word?

SPEAK: how often do you speak the word?

LISTEN: how often do you listen to the word?

In this design, we split the judgements between character-based (WRITE and READ) and voice-based (SPEAK and LISTEN) judgements and between production (WRITE and SPEAK) and reception (READ and LISTEN) judgements. The participants gave five ratings for each factor, ranging from 5 (well known/often used) to 1 (little known/rarely used).

The rating data were collected not in person but on a crowdsourcing platform. We used 'Yahoo! crowdsourcing'; 3,392 participants judged the word familiarity rates. The participants checked a stimulus word and answered rating scores for KNOW, WRITE, READ, SPEAK, and READ; at least 16 answers were collected for each word. The data were gathered on November, 2018. The data collection, which cost 1,455,494 yen, was completed within two weeks.

3.2 Model

The collected rating data is biased due to the use of the particular subject participants, which necessitates that statistical methods should be used to resolve the biases. We used a Bayesian linear mixed model to measure the ratings. The graphical model used to estimate the ratings is shown in Figure 3: N_{word} is the number of words, and N_{subj} is the number of participants; Index $i : 1 \dots N_{word}$ is the index of words, and index $j : 1 \dots N_{subj}$ is the index of participants; and $y^{(i)(j)}$ is the rating of KNOW, WRITE, READ, SPEAK, LISTEN, in which y is generated by a Normal distribution with $\mu^{(i)(j)}$ and σ, as follows:

$$y^{(i)(j)} \sim Normal(\mu^{(i)(j)}, \sigma).$$

Here, the σ is a hyper-parameter of the standard deviation, and $\mu^{(i)(j)}$ is a linear formula of slopes $\gamma^{(i)}_{subj}$, slopes $\gamma^{(i)}_{word}$ and an intercept α:

$$\mu^{(i)(j)=\alpha+\gamma^{(i)}_{word}+\gamma^{(j)}_{subj}}.$$

The slopes are modelled by a Normal distribution with the hyper-parameters of μ_{word}, σ_{word}, μ_{subj}, σ_{subj} (means and standard deviations):

$$\gamma^{(i)}_{word} \sim Normal(\mu_{word}, \sigma_{word}),$$

$$\gamma^{(j)}_{subj} \sim Normal(\mu_{subj}, \sigma_{subj}).$$

The word familiarity rates are composed by $\gamma^{(i)}_{word}$. On the other hand, the biases of subject participants are modelled by $\gamma^{(j)}_{subj}$. We set the means μ_{word} and μ_{subj} as 0.0 to make the average 0.0; we also set the standard deviations σ_{word} and σ_{subj} as 1.0. We used R and Stan to model the data. We set an iteration at 5,000 $\times$ 4 chains with an initial warm-up of 100 iterations.

4 Data Analysis

This section describes the qualitative evaluation of the estimated word familiarity rate data. To evaluate the data, we first reviewed the distribution of the five perspectives and the biases of the subject participants. Second, we confirmed the top and bottom 10 words of the estimated values. Third, we also reviewed the top and bottom 10 categories by the WLSP's second semantic category for the estimated values.

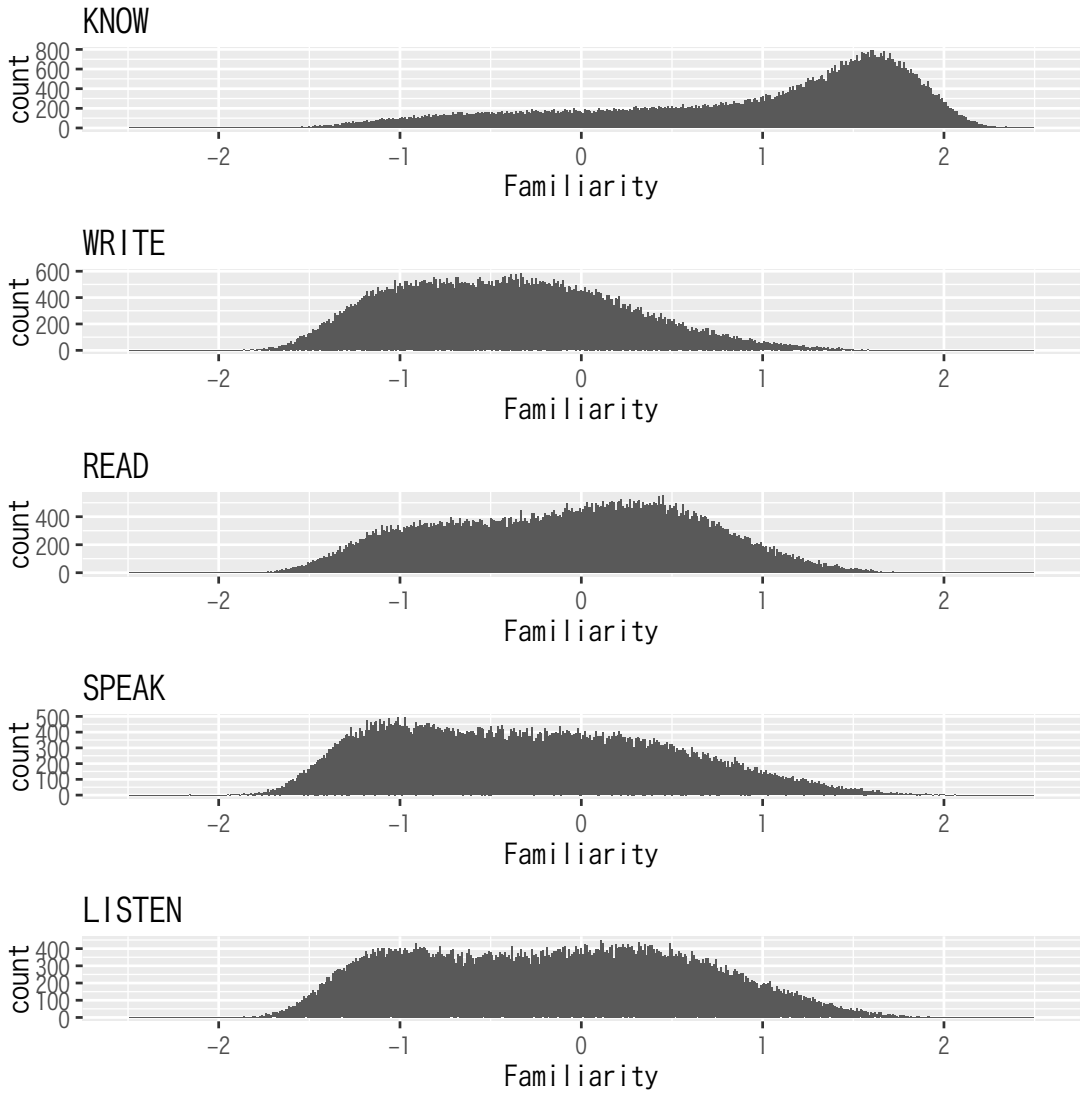


Figure 1: Estimated Familiarities ($\gamma^{(i)}_{word}$): The Distribution of the Five Perspectives

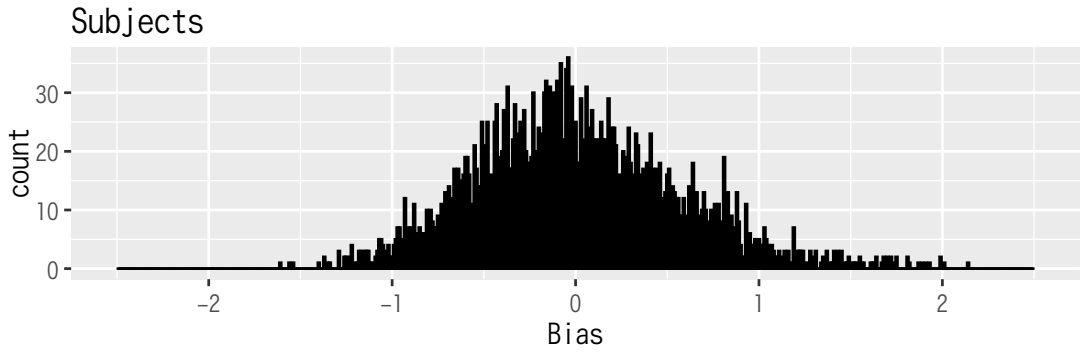
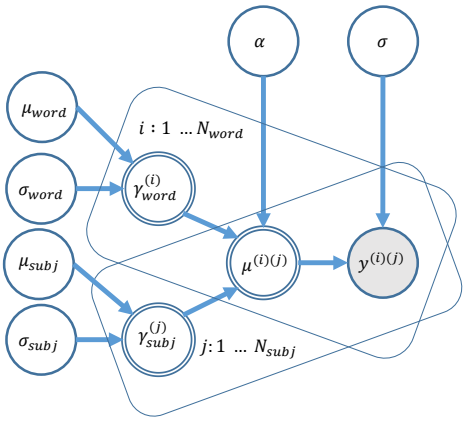


Figure 2: Estimated Biases for the Subject Participants ($\gamma^{(j)}_{subj}$)

4.1 Distributions

Figure 1 displays the histogram of the estimated familiarities. The x-axis specifies the word famil-



$$\gamma_{word}^{(i)} \sim Normal(\mu_{word}, \sigma_{word})$$
$$\gamma_{subj}^{(j)} \sim Normal(\mu_{subj}, \sigma_{subj})$$

$$\mu^{(i)(j)} = \alpha + \gamma_{word}^{(i)} + \gamma_{subj}^{(j)}$$
$$y^{(i)(j)} \sim Normal(\mu^{(i)(j)}, \sigma)$$

Figure 3: Graphical model for the Ratings

iarity rating $\gamma_{word}^{(i)}$, and the y-axis specifies the frequencies. The five perspectives are distinguished in the histogram with different colours. As illustrated in Figure 1, KNOW has a higher familiarity rating than the other perspectives, since it is the most fundamental perspective. The character-based perspectives (WRITE and READ) had lower familiarity ratings than the voice-based perspectives (SPEAK and LISTEN). Furthermore, the production perspectives (WRITE and SPEAK) had lower familiarity ratings than the reception perspectives (READ and LISTEN).

Figure 2 displays the histogram of the estimated subject participant biases. The x-axis specifies the estimated subject participant biases $\gamma_{subj}^{(j)}$, and the y-axis specifies the frequencies. The subject participant biases are modelled with standard normal distributions. We should introduce other distributions for the biases in our future work. We did attempt to use other distributions in the model; however, only the standard normal distribution converged. In future work, we will increase the amount of rating data and again attempt to use other distributions.

4.2 Evaluation by Words

In this section, we describe the top (KNOWN) and bottom (UNKNOWN) 10 words for several perspectives.

4.2.1 Known vs. Unknown

First, we reviewed KNOW, which is the most fundamental perspective.

Tables 2 and 3 display the top 10 known and unknown words for the perspective KNOW, respectively. The known words are ones that tend to be used in daily social life, while the unknown words are never or rarely used in Japan. Though we also analysed the other perspectives {WRITE, READ, SPEAK, LISTEN}, we omitted tables for the remaining four perspectives due to the limited space.

Table 2: The Top 10 Known Words (KNOW)

Words		KNOW
全員	all	2.44
恋人	lover	2.44
翌朝（よくあさ）	next morning	2.44
退社する	leave the office	2.38
再会	reunion	2.38
本社	headquarters	2.38
入社	enter a company	2.37
人見知りする	timid	2.36
持ち帰る	take away	2.36
ストロー	straw	2.36

Table 3: The Top 10 Unknown Words

Words		KNOW
うずみひ	embeded gutter	-1.86
玉章（たまずさ）	letter	-1.86
御稜威（みいつ）	authority	-1.85
繞（にょう）	kanji radical	-1.85
執掌（おうしょう）する	being busy with	-1.84
スフ	staple fibre	-1.82
驍名	valor	-1.79
笈摺（おいずり）	sleeveless overgarment worn by pilgrims	-1.79
宇内（うだい）	the whole world	-1.76
賢察	hypothesise	-1.75

4.2.2 Character-based vs. Voice-based

Next, we surveyed the difference between the character-based (WRITE/READ) and voice-based (SPEAK/LISTEN) results by evaluating the values

Table 4: Character-based Biased Words

Words		Ch-Vo
上記	the abovementioned	3.88
追伸	postscript	2.65
前述する	mentioned earlier	2.42
後述	mention later	2.35
記	description	2.30
前略	dispensing with the preliminaries	2.29
在中	enclosed	2.18
アンパサンド［&］	ampersand	2.17
句読点	punctuation	2.12
下記	the undermentioned	2.00

Ch-Vo: WRITE + READ - SPEAK - LISTEN

Table 5: Voice-based Biased Words

Words		Ch-Vo
レジ袋	shopping bag	-3.07
先っちょ	tip	-2.65
ちょろまかす	embezzle	-2.59
バイバイ	bye bye	-2.59
ヨーグルト	yoghurt	-2.52
ドライヤー	dryer	-2.47
まんま［その～］	as it is	-2.46
それではまた	see you again	-2.42
鼻水	mucus	-2.42
どっこいしょ	oof!	-2.41

Ch-Vo: WRITE + READ - SPEAK - LISTEN

Table 6: Production Biased Words

Words		P-R
毛管	capillary tube	0.76
物心（ぶっしん）	matterand mind	0.73
消却する	erase	0.73
絆創膏	adhesive tape	0.72
ふたとせ	two years	0.71
揚げなべ	deep fryers	0.71
吟詠する	sing a song	0.71
だるい	feel weary	0.69
上辺（うわべ）	outward appearance	0.68
幽寂	sequestered	0.66

P-R: WRITE + SPEAK - READ - LISTEN

Table 7: Reception Biased Words

Word		P-R
送検する	commit someone to trial	-2.93
右翼	right wing	-2.71
書類送検	filing charges	-2.69
巡業する	take a provincial tour	-2.59
西郷隆盛	Takamori Saigo	-2.52
殺害（さつがい・せつがい）	murder	-2.52
革命児	revolutionary	-2.48
護衛する	guard	-2.47
識者	well-informed people	-2.42
再審	retrial	-2.41

P-R: WRITE + SPEAK - READ - LISTEN

for (WRITE + READ - SPEAK - LISTEN). The difference between character-based (WRITE and READ) and voice-based (SPEAK and LISTEN) stimuli can be observed in the '*Nihongo no goi tokusei*' database. Here, if the value is positive, the word tends to be used in written language. If the value is negative, the word tends to be used in spoken language.

Table 4 shows the positively-valued examples. These words tend to be used in written documents or letters. Punctuation-related words 'アンパサンド (ampersand)' and '句読点 (punctuation)' also appeared in the top 10 words. Table 5 shows the negatively-valued examples. These words tend to be used in conversations in daily life. The greeting 'バイバイ (bye bye)' and the interjection 'どっこいしょ (oof!)' are also observed.

4.2.3 Production vs. Reception

We surveyed the difference between the production (WRITE/SPEAK) and reception (READ/LISTEN) results and evaluated the (WRITE + SPEAK - READ - LISTEN) values. This approach is unique because no existing research has evaluated these perspectives.

The difference between production and reception thus seems to reflect whether or not the word is used in both mass media and in normal speech. Table 6 shows the production biased words, which tend to be technical terms. Some of the subject participants ' work histories (e.g. in the medical or music fields) explain certain words in Table 6, such as '毛管 (capillary tube)' and '絆創膏 (adhesive tape)' or traditional music '吟詠する (sing a song)'. Table 7 shows the reception biased words, and the negative words ('殺害 (murder)' and '書類送検 (filling charges)') are confirmed. The word '西郷隆盛 (Takamori Saigo)' also appears as a reception biased word in Table 6, which is the main character in a TV drama.

4.3 Evaluation by WLSP categories

This section presents our evaluation of the WLSP categories. We evaluated the results using the second level of the semantic category in the WLSP, which includes two fractional digits to the right of the radix point (as explained in section 2.1). We also present the most and least familiar words in the same WLSP categories.

Table 8: The Top 10 Known Categories

	Category		KNOW
3.53	相-自然-生物	Modifier-Nature-Creature	1.41
3.17	相-関係-空間	Modifier-Relation-Space	1.41
2.10	用-関係-真偽	Verb-Relation-Truth	1.35
3.56	相-自然-身体	Modifier-Nature-Body	1.34
2.56	用-自然-身体	Verb-Nature-Body	1.32
2.14	用-関係-力	Verb-Relation-Power	1.32
3.35	相-活動-交わり	Relation-Action-Inter Course	1.32
4.32	他–呼び掛け	Other-Vocative	1.31
4.31	他–判断	Other-Judgement	1.29
3.57	相-自然-生命	Modifier-Nature-Life	1.26

Table 9: The Top 10 Unknown Categories

	Category		KNOW
3.52	相-自然-天地	Modifier-Nature-World	0.13
1.54	体-自然-植物	Noun-Nature-Botanical	0.40
1.55	体-自然-動物	Noun-Nature-Animal	0.64
1.31	体-活動-言語	Noun-Action-Language	0.66
1.23	体-主体-人物	Noun-Subject-Person	0.67
1.42	体-生産物-衣料	Noun-Product-Garments	0.68
1.52	体-自然-天地	Noun-Nature-World	0.70
1.32	体-活動-芸術	Noun-Action-Art	0.71
4.50	他-動物の鳴き声	Other-Animal Call	0.72
1.51	体-自然-物質	Noun-Nature-Material	0.76

4.3.1 Known vs. Unknown

Tables 8 and 9 display the top 10 known and unknown word categories based on the perspective KNOW, respectively. As illustrated in Tables 8 and 9, the known words tend to be modifiers or verbs, while the unknown words tend to be nouns. The most well-known category is 3.53 (相-自然-生物: Modifier-Nature-Creature), which includes gender-related words such as '女性的 (feminine)' (KNOW=1.81) and '男性的 (masculine)' (1.71). The least known category is 3.52 (相-自然-天地: Modifier-Nature-World), which includes rarely used words such as '蕭条 (bleak)' (-1.46) and '巍巍 (big and high)' (-1.35).

4.3.2 Character-based vs. Voice-based

Figures 10 and 11 display the results for the character-based biased and voice-based biased categories, respectively. As shown in these tables, the nominal action and subject categories tend to be character-based biased, whereas the vocative, interjection, modifiers, and animal call categories tend to be voice-based biased. The highest-valued character-based category is 1.31 (体-活動-言語: Noun-Action-Language), which includes epistolary words such as '上記 (aforementioned)' (WRITE+READ-SPEAK-LISTEN=3.87) and '追伸 (p.s.)' (2.65). The lowest valued voice-based biased category is 4.32 (他-呼びかけ: Other-Vocative), which includes 'もしもし (hello on phone)' (-1.75).

Table 10: Character-based Biased Categories

	Category		Ch-Vo
1.31	体-活動-言語	Noun-Action-Language	0.13
1.32	体-活動-芸術	Noun-Action-Art	0.11
1.25	体-主体-公私	Noun-Subject-Public Private	0.11
1.23	体-主体-人物	Noun-Subject-Person	0.10
1.27	体-主体-機関	Noun-Subject-Organisation	0.10
1.52	体-自然-天地	Noun-Nature-World	0.09
1.36	体-活動-待遇	Noun-Action-Treatment	0.08
2.31	用-活動-言語	Verb-Action-Language	0.07
1.53	体-自然-生物	Noun-Nature-Creature	0.07
3.52	相-自然-天地	Modifier-Nature-World	0.07

Ch-Vo: WRITE + READ - SPEAK - LISTEN

Table 11: Voice-based Biased Categories

	Category		Ch-Vo
4.32	他–呼び掛け	Other-Vocative	-0.59
4.30	他–感動	Other-Interjection	-0.53
3.56	相-自然-身体	Modifier-Nature-Body	-0.44
2.56	用-自然-身体	Verb-Nature-Body	-0.43
3.51	相-自然-物質	Modifier-Nature-Material	-0.42
3.18	相-関係-形	Modifier-Relation-Form	-0.33
3.50	相-自然-自然	Modifier-Nature-Nature	-0.30
3.57	相-自然-生命	Modifier-Nature-Creature	-0.29
4.50	他-動物の鳴き声	Other-Animal Call	-0.29
1.43	体-生産物-食料	Noun-Product-Food	-0.28

Ch-Vo: WRITE + READ - SPEAK - LISTEN

4.3.3 Production vs. Reception

Tables 12 and 13 display the results for the production biased and reception biased categories, respectively. Generally, the reception values (READ, LISTEN) tend to be larger than the production values (WRITE, SPEAK). Therefore, the

Table 12: Production Biased Categories

	Category		P-R
4.50	他-動物の鳴き声	Other-Animal Call	-0.26
2.10	用-関係-真偽	Verb-Relation-Truth	-0.27
4.30	他-感動	Other-Interjection	-0.29
1.54	体-自然-植物	Noun-Nature-Botanical	-0.30
4.32	他-呼び掛け	Other-Vocative	-0.30
3.52	相-自然-天地	Modifier-Nature-World	-0.32
4.11	他-接続	Other-Conjunction	-0.35
1.42	体-生産物-衣料	Noun-Product-Garments	-0.35
1.55	体-自然-動物	Noun-Nature-Animal	-0.35
4.31	他-判断	Other-Judgement	-0.36

P-R: WRITE + SPEAK - READ - LISTEN

Table 13: Reception Biased Categories

	Category		P-R
1.27	体-主体-機関	Noun-Subject-Organization	-0.62
1.36	体-活動-待遇	Noun-Action-Treatment	-0.56
1.35	体-活動-交わり	Noun-Action-Intercourse	-0.55
1.53	体-自然-生物	Noun-Nature-Creature	-0.54
3.17	相-関係-空間	Modifier-Relation-Space	-0.54
1.24	体-主体-成員	Noun-Subject-Member	-0.54
2.35	用-活動-交わり	Verb-Action-Inter Course	-0.53
2.36	用-活動-待遇	Verb-Action-Treatment	-0.53
2.34	用-活動-行為	Verb-Action-Behaviour	-0.52
3.14	相-関係-力	Verb-Relation-Power	-0.52

P-R: WRITE + SPEAK - READ - LISTEN

values for Pro-Rec (WRITE + SPEAK - READ - LISTEN) become negative, even for the production biased categories. The syntactic categories (excluding nouns, verbs, and modifiers) are production biased such as the animal call, interjection, vocative, and conjunction categories. The other production biased category is 4.50 (他-動物の鳴き声: Other-Animal Call), which includes words such as 'げろげろ (croak)' (WRITE+SPEAK-READ-LISTEN=0.45) and 'かーかー (croak)' (0.23). The reception biased words refer to the vocabulary used on the news or in TV show such as nominal organisation, treatment, or intercourse. The reception biased category with the highest ranking is 1.27 (体-主体-機関: Noun-Subject-Organization), which includes words such as '厚生労働省 (Ministry of Health, Labour, and Welfare)' (-2.23) and '金融庁 (Financial Services Agency)' (-2.18).

4.4 Discussions

In this paper, we presented the word familiarity rating tendencies based on a crowdsourced study. The character-based (WRITE and READ) /voice-based (SPEAK and LISTEN) contrasting results confirm the findings in *Nihongo no goi tokusei*; however, in our data, we uniquely observe the contrast between the production and reception categories.

However, we still face the issue of normalising the ratings. This study ' s proposed method, in which the mean and standard deviation are set to 0.0 and 1.0, respectively, is sufficient when rating relative values or when arranging ratings in a certain order. We also calculated the ratings with $\gamma_{word}^{(i)} + \mu_{subj} + \alpha$; with this calculation, the ratings can be ranged from 1.0 to 5.0, excluding outliers. Though the normalization of ratings should be determined by the rating method used, calculating the value $\gamma_{word}^{(i)}$ is sufficient for most uses.

5 Conclusions

We have presented a Japanese word familiarity rate database for entries in the WLSP. To do so, we used crowdsourcing to explore the word familiarity ratings in terms of five perspectives: KNOW, WRITE, READ, SPEAK, and LISTEN. A Bayesian linear mixed model was utilised to estimate the ratings. The data [3] and code[4] are publicly available. Our future work on this topic is as follows. In this paper, we modelled the word familiarity rates and the subject participant biases with the standard normal distribution. While we did attempt to model the rates and biases with other distributions, the MCMC estimation did not converge. In the future, we hope to perform the survey on a yearly basis (to enlarge the data size) in order to model other distributions. We will also enhance the target word list to include UniDic entries for content words. In addition, we plan to create a morphological analyser, which will extract the word familiarity rates.

Acknowledgments

This work was supported by JSPS KAKENHI Grants Number 17H00917, 18H05521, 18K18519, 19K00591, 19K00655 and a project of the Center for Corpus Development, NINJAL.

[3] https://cradle.ninjal.ac.jp/

[4] https://github.com/masayu-a/WLSP-familiarity

References

Shigeaki Amano and Tadahisa Kondo, editors. 1999. *Nihongo no goi tokusei (Lexical properties of Japanese)*. Sanseido, Tokyo.

Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. 2014. Archiving and analysing techniques of the ultra-large-scale web-based corpus project of ninjal, japan. *Alexandria*, 1–2:129–148.

Kokuritsu_Kokugo_Kenkyusho. 1964. *Bunrui goihyo (Word List by Sementic Principles)*. Shuei Shuppan, Tokyo.

Kokuritsu_Kokugo_Kenkyusho. 2004. *Bunrui goihyo zouho kaitei-ban (Word List by Sementic Principles, Revised and Enlarged Edition)*. Dainippon Tosho, Tokyo.

Asuko Kondo, Makiro Tanaka, and Masayuki Asahara. 2018. Alignment table between unidic and 'word list by semantic principles'. In *Proceedings of The Eighth Conference of Japanese Association for Digital Humanities (JADH2018)*, pages 125–128.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of japanese. In *Proceedings of LREC-2000, (Second International Conference on Language Resources and Evaluation)*, volume 2, pages 947–952.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written japanese. *Language Resources and Evaluation*, 48(2):345–371.

Tanner Sorensen, Sven Hohenstein, and Shravan Vasishth. 2016. Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, 12(3):175–200.