Word Familiarity Rate Estimation for Japanese Functional Words Using a Bayesian Linear Mixed Model

Bocheng Chen and **Masayuki Asahara**
Graduate Institute for Advanced Studies, SOKENDAI, Japan
National Institute for Japanese Language and Linguistics, Japan
gs20233502@ninjal.ac.jp and masayu-a@ninjal.ac.jp

Abstract

This paper presents research on word familiarity rate estimation using a Japanese functional word lexicon 'Tsutsuji'. We collected rating information on 6,396 surface forms (6,870 words) in the lexicon using Yahoo! crowdsourcing. We asked 3,566 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 50 subject participants. We used Bayesian linear mixed models to estimate the word familiarity rates.

1 Introduction

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. Additionally, lexicons encompass not only linguistic information but also real-world language usage in daily life. There are two methods for database construction based on the actual usage of words. One is frequency-based data derived from corpora, and the other is word-familiarity data obtained through questionnaire surveys. In the case of Japanese, one such language resource is the 'Word Familiarity Rate' of the Nippon Telegraph and Telephone Corporation (NTT) (Amano and Kondo, 1999). Recently, a new version of this resource was compiled by NTT (Fujita and Kobayashi, 2020; NTT, 2021). The National Institute for Japanese Language and Linguistics (NINJAL) also developed a word familiarity rate database(Asahara, 2019) based on the 'Word List by Semantic Principles'(Kokuritsu_Kokugo_Kenkyusho, 2004).

Regarding the word familiarity rate in Japanese, to date, most research has focused on content words, with limited studies on functional words. In Japanese, a functional expression database has been constructed called 'Tsutsuji', and this study conducted surveys based on it following the work of Asahara (2019). The database includes functional words, such as multiword expressions, contracted forms, and inflections, and contains both frequently used and rarely used forms. Through this survey, it is possible to identify obsolete functional expressions. We asked participants to rate their familiarity with the words from five perspectives: KNOW, WRITE, READ, SPEAK, and LISTEN. The quality of the results gathered by crowdsourcing may be lower than that of the results collected in a controlled experiment; however, the cost of constructing a crowdsourced study is lower than that of conducting an experiment. We used a Bayesian linear mixed model to alleviate noise in the data. Furthermore, by gathering information from the perspectives of WRITE, READ, SPEAK, and LISTEN, it is possible to determine whether the usage falls under written or spoken language, thus revealing the register.

This study makes the following contributions to the literature.

- We compiled a word familiarity rate database for functional word lexicons, including multiword expressions and contracted forms.
- We used crowdsourcing via human subject participants to explore word ratings, and introduced a Bayesian linear mixed model for this type of rate modelling.
- We introduced the contrast between character-based (WRITE, READ) and voice-based (SPEAK, LISTEN) perspectives. We contribute to the literature also by 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 Familiarity Rate' in Japanese and the functional expression lexicon 'Tsutsuji'. Section 3 presents

	Amano and Kondo (1999)	Fujita and Kobayashi (2020)	Asahara (2019)
Tokens	76,945 (32,443)	163,017	100,830
Rating	1-7	1-7	1-5
Base Lexicon	Shinmeikai Kokugojiten 4th (Gakken Kokugo Daijiten 2nd)	Amano and Kondo (1999), Youjigoihattatsu DB, BCCWJ, NTT Ehon DB	Word List by Semantic Principles

Table 1: Previous work: Japanese word familiarity databases

the methodology used to develop the word familiarity ratings, namely, crowdsourcing and a Bayesian linear mixed model. Section 4 evaluates the results, and Section 5 presents the conclusions and discusses future research.

2 Related Work

First, we provide three examples of studies of word familiarity rates in Japanese. Table 1 lists the previous work on Japanese word familiarity databases. Amano and Kondo (1999) conducted a study of word familiarity rates in Japanese utilising the 'Word Familiarity Rate' dataset developed by NTT. The first version of the database included 76,945 tokens from Shinmeikai Kokugojiten (4th Edition). The expanded version of the database includes 32,443 tokens from the Gakken Kokugo Daijiten (2nd Edition). (Fujita and Kobayashi, 2020) expanded on the previous research by NTT and developed an updated version of the word familiarity rate database. This new database expands the coverage of word familiarity assessments in Japanese. The database includes 163,017 tokens from Amano and Kondo (1999), Youjigoihattatsu DB (Kobayashi et al.), the Balanced Corpus of Contemporary Written Japanese (Maekawa et al., 2014), and NTT Ehon DB (Fujita et al., 2018). These two are based on familiarity ratings between 1 and 7. (Asahara, 2019) constructed a word familiarity database based on the 'Word List by Semantic Principles' (Kokuritsu_Kokugo_Kenkyusho, 2004). He introduced the five perspectives of KNOW, WRITE, READ, SPEAK, and LISTEN for word familiarity rates. The database is based on a rating between 1 and 5 for familiarity, and includes 100,830 tokens. These databases are used primarily for Japanese content words. To date, there is no functional expression database containing word familiarity ratings.

Second, we explain the 'Tsutsuji' functional expression database (Matsuyoshi et al., 2006). Tsutsuji uses a hierarchy with nine abstraction levels (Table 2): the root node is a dummy node that governs all entries, the node in the first level is a lemma in the lexicon, and the leaf node corresponds to the surface form of a functional expression. This hierarchy also provides a method to systematically generate different surface forms. They compiled a dictionary with 341 headwords (Level 1) and 16,801 surface forms (Level 9) covering almost all major functional expressions. However, this lexicon contains obsolete or outdated expressions. We explored word familiarity ratings for the functional expressions.

Level		count
Level 1	Lemma	341
Level 2	Senses	435
Level 3	Deviation	555
Level 4	Alternation	774
Level 5	Phonological changes	1,187
Level 6	Insertion of focus particle	1,810
Level 7	Conjugation	6,870
Level 8	Insertion of honorific	9,722
Level 9	Orthographic variation	16,801

Table 2: Hierarchy of 'Tsutsuji'-1.1

3 Methodology

3.1 Rating Information Collection

We present our methodology for constructing a word familiarity rate lexicon of functional words, including multiword expressions.

We use 'Tsutsuji' (Matsuyoshi et al., 2006) as the base lexicon. We explored Level 7 of the lexicon with 6,870 entries. Because the data included polysemous words, the statistical model was constructed using 6,396 surface forms.

Figure 1 shows an example of a survey form. The following five perspectives were collected:

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?

Figure 1: An Example Survey Form

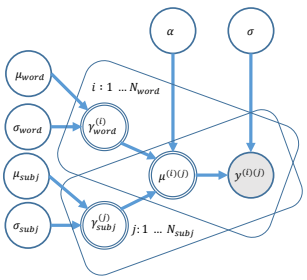
LISTEN: how often do you listen to the word?

In this design, we split judgements into character-based judgments (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’; 3,566 participants judged the word familiarity rates. The participants checked a stimulus word and provided rating scores for KNOW, WRITE, READ, SPEAK, and READ; at least 50 answers were collected for each word. Data were collected on 29 December 2022. Data collection, which cost 377,850 yen, was completed within six hours.

3.2 Model

The collected rating data were biased because of the use of particular subject participants, which necessitates the use of statistical methods to resolve biases. We used a Bayesian linear mixed model to measure the ratings. The graphical model used to estimate the ratings is shown in Figure 2, where N_{word} is the number of words, and N_{subj} is the number of participants. Index $i : 1 \ldots N_{word}$ is



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

Figure 2: Graphical Model for the Ratings

the index of words, index $j : 1 \ldots N_{subj}$ is the index of the participants, and $y^{(i)(j)}$ is the rating of KNOW, WRITE, READ, SPEAK, LISTEN, where 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, σ is a hyperparameter 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 were modelled by a normal distribution with hyperparameters 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 comprised $\gamma^{(i)}_{word}$. On the other hand, the biases of subject participants are modelled by $\gamma^{(j)}_{subj}$. We set the means μ_{word} and μ_{subj} to 0.0 to make the average 0.0; we also set the standard deviations σ_{word} and σ_{subj} to 1.0 and 0.5, respectively. R and Stan software were used to model the data. We set an iterations at 500 $\times$ 3 chains, with an initial warm-up of 50 iterations. Thus, the model converged.

4 Data Analysis

This section describes a qualitative evaluation of the estimated word familiarity rate data. To evaluate the data, we first reviewed the distribution of the participants’ five perspectives and biases. Second, we confirmed the top and bottom 10 words of the estimated values.

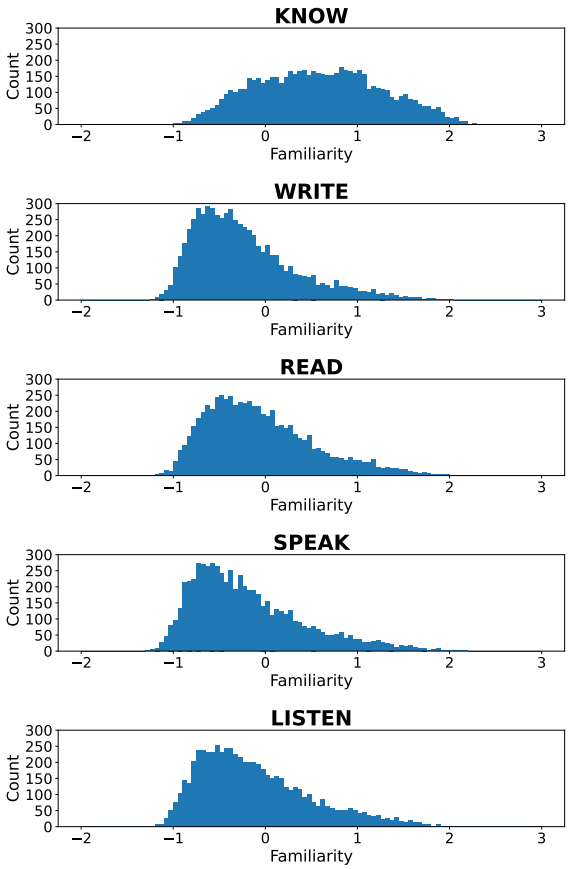


Figure 3: Estimated Familiarities ($\gamma^{(i)}_{word}$): Distribution of the Five Perspectives

4.1 Distributions

Figure 3 shows a histogram of the estimated familiarities. The x-axis specifies the word familiarity rating $\gamma^{(i)}_{word}$ 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 because it is the most fundamental perspective.

In a previous study (Asahara, 2019), the production perspectives (WRITE and SPEAK) of content words had lower familiarity ratings than the reception perspectives (READ and LISTEN). In contrast, no significant differences were observed in functional words.

4.2 Evaluation by Words

Tables 3 and 4 display the top 10 known and unknown words for the perspective KNOW, respectively. Known words tend to be used in daily social life, whereas unknown words are rarely or never used in Japan. Although we also analysed the other perspectives {WRITE, READ, SPEAK, LISTEN}, we put the tables of remaining four perspectives in Appendix A.1. By combining the four perspectives of WRITE + READ - SPEAK - LISTEN, we can estimate whether the words are character-based (positive) or voice-based (negative), as shown in Appendix A.2.

Table 3: The Top 10 Known Words (KNOW)

Words	KNOW
nado (and so on; etc.)	2.312
rashii (seem; appear)	2.306
datte-syōganai (it can't be helped)	2.299
datte (because)	2.266
nakutewa-naranai (must; should)	2.253
kudasai (please...)	2.253
to-ieba (speaking of)	2.244
mitai (seems like; looks like)	2.217
nakereba-naranai (must; have to)	2.211
de-aru (be)	2.199

Table 4: The Top 10 Unknown Words (KNOW)

Words	KNOW
tesaeko	-1.063
kotonomiikenou	-1.032
utosurasei	-1.030
chimawa	-1.012
zunishikaoka	-0.987
utomosei	-0.979
utosurashiro	-0.971
jimawa	-0.954
utomoshi	-0.936
kotonomiikenaki	-0.921

5 Conclusions

We present a functional word familiarity rate database for entries in Tsutsuji. To do so, we used crowdsourcing to explore the word familiarity ratings from five perspectives: KNOW, WRITE, READ, SPEAK, and LISTEN. A Bayesian linear mixed model was used to estimate the ratings. These four perspectives can be combined into WRITE + READ - SPEAK – LISTEN, where a positive value indicates a character-based word ((WRITE + READ) > (SPEAK - LISTEN)), and a negative value indicates a voice-based word ((WRITE + READ) < (SPEAK + LISTEN)).

The data and code[1] are publicly available.

Acknowledgements

We would like to express out gratitude for the support and collaboration provided by the National Institute for Japanese Language and Linguistics

[1] https://github.com/masayu-a/TSUTSUJI-familiarity

(NINJAL) in the collaborative research project, 'Evidence-based Theoretical and Typological Linguistics.' This research is partially supported by JSPS KAKEN JP22H00663.

References

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

Masayuki Asahara. 2019. Word familiarity rate estimation using a Bayesian linear mixed model. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 6–14, Hong Kong. Association for Computational Linguistics.

Sanae Fujita and Tessei Kobayashi. 2020. Tangoshinmitsudo-no saichousa-to kako-no de-ta-tono hikaku. In *Gengoshorigakkai dai26kai nenjitaikai happyouronbunshu*.

Sanae Fujita, Yuko Okumura, Tessei Kobayashi, and Takashi Hattori. 2018. Ehon-to youjimuke-no hatsuwa-ni shutsugensuru go-no tayouseihikaku. In *Gengoshorigakkai dai24kai nenjitaikai happyouronbunshu*.

Tessei Kobayashi, Yuko Okumura, and Yasuhiro Minami. Correcting data on child vocabulary development by vocabulary-checklist application. In *IEICE technical report*, volume 115, page 1.

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

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.

Suguru Matsuyoshi, Satoshi Sato, and Takehito Utsuro. 2006. Compilation of a dictionary of japanese functional expressions with hierarchical organization. In *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, pages 395–402, Berlin, Heidelberg. Springer Berlin Heidelberg.

NTT, editor. 2021. *NTT goi de-tabe-su*. NTT Insatsu, Tokyo.

A Evaluation by Words with Other Perspectives

A.1 The Top 10 words in WRITE, READ, SPEAK, and LISTEN

Table 5: The Top 10 Words (WRITE)

Words	WRITE
demo (but)	2.004
rashii (seem; appear)	2.003
nado (and so on; etc.)	1.996
kudasai (please...)	1.988
wo	1.987
amari (not much)	1.923
kedo (but)	1.918
yori (than; from)	1.897
gurai (or so; about; approximately)	1.880
ni-tsuite (about)	1.856

Table 6: The Top 10 Words (READ)

Words	READ
kudasai (please...)	2.118
demo (but)	2.047
nado (and so on; etc.)	2.012
wo	2.006
de-aru (be)	1.994
to-ieba (speaking of)	1.987
amari (not much)	1.969
tameni (for)	1.952
ni-tsuite (about)	1.931
yori (than; from)	1.930

Table 7: The Top 10 Words (SPEAK)

Words	SPEAK
rashii (seem; appear)	2.183
kudasai (please...)	2.162
kedo (but)	2.113
demo (but)	2.086
wo	2.064
ii (good)	2.042
to-ieba (speaking of)	2.006
kamo (maybe; may)	1.992
hazu (should; supposed to)	1.985
gurai (or so; about; approximately)	1.917

Table 8: The Top 10 Words (LISTEN)

Words	LISTEN
kudasai (please...)	2.196
kedo (but)	2.159
demo (but)	2.123
rashii (seem; appear)	2.105
mitai (seems like; looks like)	2.074
ii (good)	2.053
wo	2.050
to-ieba (speaking of)	2.027
nado (and so on; etc.)	1.969
datte (because)	1.963

Tables 5, 6, 7, and 8 list the top 10 words from the perspectives of WRITE, READ, SPEAK, and LISTEN.

A.2 Character-based vs. Voice-based (WRITE+READ-SPEAK-LISTEN)

Table 9: Character-based Words

Words	Ch-Vo
ni-okeru (in; at)	1.180
de-aru (be)	1.166
no-gotoku (as; as if; like)	1.145
orikara (just now; recently)	1.086
wo-hajime-toshita (including)	1.074
kotonimonare	0.952
dewa-naranu (it cannot be)	0.946
te-kudasaranu-darōka (won't you help me)	0.939
karatote (just because; not necessarily)	0.933
ni-saishite (on the occasion of)	0.906

Ch-Vo: WRITE + READ - SPEAK - LISTEN

Table 10: Voice-based Words

Words	Ch-Vo
dakke (is it)	-1.734
kotonisurya	-1.356
teka (anyway; by the way)	-1.272
kamoshiren (maybe)	-1.121
teittatte	-1.089
tatte-shōganai (it can't be helped)	-1.050
ja-dame (must not; not good)	-1.044
tara-ii (it would be good)	-1.037
cha-ikan (must not; not good)	-1.035
damon (it's because)	-1.032

Ch-Vo: WRITE + READ - SPEAK - LISTEN

Next, we surveyed the difference between the character-based (WRITE/READ) and voice-based (SPEAK/LISTEN) results by evaluating the values for (WRITE + READ - SPEAK - LISTEN). If the value is positive, the word tends to be used in the written language. If the value is negative, the word tends to be used in the spoken language. Whereas the character-based words in Table 9 include literary expressions, the voice-based words in Table 10 include colloquial expressions such as contracted forms.