Clarifying the Role of Psychological Factors in Language Acquisition: A Psycholinguistic Lexical Ratings Dataset

Wanwan Zheng
Humanity Center for Anthropocenic Actors and Agency, Graduate School of Humanities
Furo-cho, Chikusa, Nagoya 464-8601, Japan
zheng@nagoya-u.jp

Abstract

Lexical acquisition extends beyond the learning of surface-level word forms and encompasses underlying cognitive characteristics as well as broader processes that involve the learner. Nevertheless, in Japanese, as in many other languages, word difficulty has been characterized primarily by frequency and surface-level properties. Far less is known about the cognitive and affective dimensions that influence whether words are easier or more difficult to process. To address this gap, this study introduces a novel dataset that incorporates six psycholinguistic dimensions—familiarity, affective valence, arousal, imageability, abstractness, and understandability—collected through large-scale surveys of Japanese second language learners. Preliminary analyses of responses from 536 participants across 15 countries demonstrated that the dataset is both theoretically coherent and empirically reliable, consistent with established theories and findings while also yielding new insights into lexical processing. In addition to supporting more accurate estimations of word difficulty, the dataset provides a resource for future research by enabling systematic exploration of how lexical processing is shaped through the interaction of visual, affective, cognitive, and contextual factors.

Keywords: word difficulty, psycholinguistic dimensions, second language acquisition (Japanese), cognitive and affective processing, lexical dataset

1. Introduction

Empirical evidence has demonstrated that the accurate identification of words constitutes a fundamental prerequisite for reading comprehension (Duff et al., 2015; Killingly et al., 2025). Consistent with this view, interactive models of text comprehension developed in the 1990s highlighted the dynamic interplay between bottom-up processes (e.g., word recognition and other lower-level operations) and top-down processes (e.g., meaning construction and the integration of background knowledge) as essential components of successful reading comprehension (Kintsch, 1998; Perfetti, 1999). Moreover, vocabulary coverage not only constrains the range of thought but also serves as a robust indicator of cognitive ability (Bruce and Bell, 2022). Research with preschool children has further revealed a significant bidirectional relationship between executive functioning and vocabulary size (Schmitt et al., 2019). Therefore, vocabulary knowledge entails more than the acquisition of linguistic symbols; it also supports efficient language processing and the development of cognitive control (Bransford et al., 2000; Jucks and Paus, 2012). Consequently, the capacity to construct meaning from text, as well as to regulate and adapt to cognitive activity, can be understood as fundamentally rooted in word-level processing.

On the other hand, information about word difficulty contributes to more efficient vocabulary acquisition. However, word difficulty is determined by multiple factors, including morphological features (e.g., word length and the number of morphemic components), semantic features (e.g., polysemy and domain specificity), exposure-related variables (e.g., frequency and familiarity), imageability and concreteness (i.e., the extent to which a word elicits mental imagery), and psychological factors (e.g., affective valence, ease of cognitive processing, and memory retention). Consequently, the estimation of word difficulty remains a complex task. In current linguistic and educational research, the common practice has been the manual selection of words assigned to different difficulty levels. Such human evaluation, however, is labor-intensive and inherently restricted in scale. For this reason, mechanical and quantitative approaches to word difficulty estimation have become increasingly necessary. To develop reliable models, it is essential to identify the key determinants of word difficulty and to elucidate the magnitude of their contributions.

Among the numerous factors influencing word difficulty, frequency effect is the most widely recognized. High-frequently words are typically acquired more rapidly and processed more easily, whereas low-frequency words are more likely to be perceived as difficult. Hashimoto and Egbert (2019) reported a correlation of –0.50 between word frequency and difficulty. Extending this finding, Stewart et al. (2022) demonstrated that when a broader frequency range is examined and a logarithmic transformation is applied, the correlation strengthens to –0.80. Other usage-related factors include age of acquisition (Hiebert et al., 2019), regularity (e.g., word bigrams; Ha et al., 2024), and contex-

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 47–56
16 May 2026.

tual dependence (e.g., word co-occurrence; Ha et al., 2024). In contrast, intrinsic word-level properties include length (Cervetti et al., 2015), polysemy (Vitta et al., 2023; Canning et al., 2024), and the knowledge required for comprehension (Yamakawa and Tajima, 2024). As most of these indicators are derived from count-based values, their reliability depends heavily on the linguistic resources (corpora) used.

In addition to quantitative indicators, there are qualitative indicators that are less dependent on specific linguistic resources. These include (1) word familiarity (i.e., the perceived familiarity of words among language users; Leroy and Kauchak, 2014), (2) image-related semantic properties such as imageability and concreteness (Kumcu and Thompson, 2020), and (3) psychological factors such as affective valence and memory retention (Barrett, 2017). Unlike corpus-based frequency indicators, these indicators are less susceptible to textual distribution biases and may therefore offer a more universal estimation of word difficulty.

Word familiarity With respect to word familiarity, Zheng (2024) estimated word difficulty using five dimensions of familiarity: "knowing," "hearing," "speaking," "writing," and "reading." In general, familiarity scores tend to be higher for easier words and lower for more difficult ones. However, the analysis revealed two atypical categories: "easy but unfamiliar words" and "difficult but familiar words." When these outliers were removed through anomaly detection, the classification accuracy of beginner-, intermediate-, and advanced-level words improved from 0.66 to 0.93. Nevertheless, because the learning environments of native speakers and second language (L2) learners differ, there are words for which familiarity and difficulty do not align across groups. This discrepancy suggests that familiarity data used in this work may reflect a mixture of responses from both native speakers and L2 learners.

Image-related semantic factors With respect to image-related semantic factors, it is well established that words referring to perceptually salient and highly imageable objects or concepts are recalled more easily than those associated with less imageable ones (Marschark and Cornoldi, 1991). In this context, Kumcu and Thompson (2020) argued that low-difficulty words are typically more imageable and concrete, thereby reducing cognitive load and facilitating more accurate recall.

Affective valence and arousal With respect to affective valence, emotion has been recognized as a central concept in applied linguistics, playing a fundamental role in cognition and receiving increasing attention for its role in foreign language acquisition (Dewaele, 2015). Previous research has introduced concepts such as "emotionally enhanced memory" (Talmi et al., 2007) and "affective input enhancement" (Truscott, 2015). Moreover, Kanazawa (2021, 2023) demonstrated that emotion-based processing facilitates the retrieval of target words. Hashimoto and Egbert (2019) further incorporated affective valence—an emotional dimension associated with word difficulty—alongside with other lexical variables such as word length and contextual distinctiveness. However, linguistic resources developed on the basis of native speakers' responses are not necessarily applicable to foreign language learners, whose mental lexicons differ substantially. This discrepancy underscores the importance of developing foreign language-specific versions of such databases, particularly for subjective attributes (Kanazawa, 2021).

Taken together, these findings indicate that word difficulty cannot be explained by a single factor; rather, it must be evaluated within a broader multidimensional framework that integrates both linguistic and psycholinguistic factors. The Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle et al., 2018), developed to assess the difficulty of English words, incorporates more than 400 indicators. Because most of these indicators are statistical measures, such as frequency, they are relatively straightforward to interpret; however, they remain dependent on corpora or dictionary-based databases. In contrast, lexical properties grounded in psychological and cognitive factors (e.g., familiarity, imageability, abstractness, and affective valence) capture dimensions of word difficulty that cannot be fully explained by frequency-based statistics alone.

To address these limitations, this study constructed a dataset for learners of Japanese as a second language with the aim of enabling more precise estimation of word difficulty. The dataset incorporates six psycholinguistic features—familiarity, affective, arousal, imageability, abstractness, and understandability—thereby capturing the cognitive complexity of lexical processing from multiple perspectives.

The remainder is structured as follows. Section 2 describes the procedures used to develop the dataset. Section 3 presents the data structure both quantitatively and visually. Section 4 evaluates the effectiveness of the dataset in estimating word difficulty and reports benchmark experiments. Finally, Section 5 concludes with a summary.

2. Data Collection

2.1. Selection of Target Words

The objective of this study was to construct a comprehensive and reliable dataset for the evaluation of word difficulty. As discussed in the preceding sections, word difficulty is a multidimensional construct shaped by diverse linguistic, cognitive, and psychological factors. Accordingly, the availability of highly reliable word difficulty information of paramount importance for both dataset development and subsequent analyses.

For this purpose, the words included in the *Japanese Language Learner's Dictionary* were selected as the targeted words (Sunakawa et al., 2012). This dictionary contains 17,920 entry words derived from two major lexical resources: (1) a lexical survey of the *Balanced Corpus of Contemporary Written Japanese* (BCCWJ)—a corpus specifically designed to capture the overall characteristics of contemporary written Japanese and currently the only available balanced corpus of the language—and (2) a corpus of Japanese language textbooks, comprising electronic data from 100 commercially available textbooks ranging from beginner to advanced levels (unpublished resource). The adoption of this dictionary as the basis for target word selection provides distinct advantages. Because its vocabulary was systematically sampled from both a balanced national corpus and pedagogical materials spanning all proficiency levels, the resulting word list ensures broad representativeness across genres and registers while maintaining direct applicability to second-language learning contexts.

In detail, each entry in the dictionary is annotated with multiple layers of information, including standardized orthographic form, spelling, difficulty level, part of speech (POS), and word type (i.e., native Japanese words, Sino-Japanese words, loanwords, and hybrid formations). The difficulty of each entry word was classified into six different levels—lower/upper elementary, lower/upper intermediate, and lower/upper advanced—based on the subjective judgment of experienced Japanese language teachers. Representative examples are provided in Table 1.

2.2. Survey Design

For each word, the survey items were presented as illustrated in Figure 1. Note that the questionnaire itself was presented in Japanese. Because Japanese includes homographs with different pronunciations, which may influence perceived difficulty, the stimuli were presented in the form of "word + reading." Participants were asked to evaluate their cognitive and affective impressions of the given word from two perspectives:

1. `Familiarity` This item assessed the degree of familiarity with the target word. Participants rated their familiarity degree across four usage contexts—writing, reading, speaking, and listening—as well as their overall impression (four items in total). Responses were provided on a seven-point Likert scale ranging from *Not at all* to *Strongly think so*.

2. `Meaning and Image` This item elicited subjective evaluations of the semantic and psychological properties of the target word. Specifically, participants rated five aspects: Negativity (affective valence: pleasant—unpleasant), Calmness (arousal: calming—exciting), Ease of evoking an image (imageability: ease of eliciting a mental representation), Abstractness (degree of conceptual difficulty), and Ease of understanding (understandability: perceived difficulty of learning). These responses were also recorded on a seven-point Likert scale.

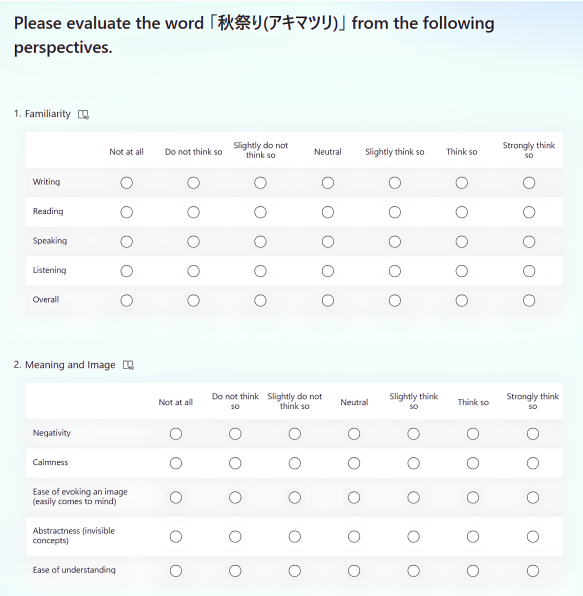

Figure 1: Survey items (translated). The questionnaire itself was presented in Japanese.

In this survey, participants were presented with a set of seven words, and two sets (14 words in total) were administered. These words were randomly selected from the vocabulary list of the applied dictionary, with the design ensuring representation from all six levels of word difficulty. Because the task did not involve correct or incorrect answers, participants' intuitive judgments were prioritized. To assess response reliability, two words were duplicated across the sets, requiring participants to evaluate the same item twice. To minimize memory and order effects, the duplicated words were

Standardized orthographic form	Spelling	Difficulty level	POS	Word type
秋祭り (autumn festival)	アキマツリ	upper intermediate	noun	native Japanese word
愛好 (fondness/enthusiasm)	アイコウ	lower advanced	noun	Sino-Japanese word
アイス (ice cream)	アイス	lower elementary	noun	loanword
空き缶 (empty can)	アキカン	lower intermediate	noun	hybrid word

Table 1: Representative examples of target words selected from the *Japanese Language Learner's Dictionary.*

inserted at fixed intervals. The presentation sequence followed the format: A (5 words) → C (2 words) → B (5 words) → C (2 words). In addition, demographic information regarding gender, age, nationality, and learning level was collected at the end of the questionnaire. For learning level, five categories ranging from beginner to advanced were provided, each accompanied by explanatory descriptions to facilitate accurate self-assessment. The question was phrased as follows:

Please indicate your level of Japanese (select the one that is closest to your level).

- Beginner (able to engage in simple greetings and self-introductions)
- Upper Beginner (able to conduct basic conversations, e.g., hobbies, schedules)
- Intermediate (able to understand slow-paced conversations and simple texts)
- Upper Intermediate (able to understand moderately complex conversations and texts, and to use Japanese in both daily life and work contexts)
- Advanced (able to understand newspapers, news, and specialized content, and to speak naturally)

2.3. Survey Administration and Screening

The questionnaire was developed using Qualtrics and administered online via URL or QR code. The study received ethical approval from the Ethics Review Committee of university (approval no. NUHM-25-17). All participants provided informed consent online prior to the start of the survey. In Japan, cooperation was requested from Nagoya University as well as 202 Japanese language schools nationwide. In China, collaboration was sought from Peking University, Jinan University, Capital Normal University, Dalian University of Foreign Languages, and Shanghai International Studies University. In addition, participation was solicited through social media platforms such as Facebook.

The survey began in Augest 2025 and is currently ongoing. The estimated completion time was approximately 10–18 minutes. At the start of the survey, participants were presented with an informed consent form describing the study, and responses were recorded only after consent was provided. As compensation, participants who submitted valid responses received an electronic gift card valued at 300 yen (15 RMB in China). To increase sample size, multiple submissions from the same participant were permitted, with a maximum of ten entries.

To ensure the validity of responses, entries with missing values were first excluded. Subsequently, for each participant, the ratings provided for the two duplicated words were compared. The Likert-scale responses, ranging from "*Not at all*" to "*Strongly think so*", were converted into numerical values on a 0–6 scale. Response consistency was then evaluated using two measures: the mean absolute error (MAE) and the exact agreement rate (Agree). The Agree index is easy to interpret; however, because the agreement rate disregards directionality, it does not differentiate between minor discrepancies (e.g., ± 1) and major discrepancies (e.g., ± 3 or more). Therefore, it was employed in combination with MAE, which captures the magnitude continuous error.

For each duplicated word $i = 1, 2$, numerical evaluations of the items $j = 1, \ldots, 10$ are denoted as a_{ij} (first rating) and b_{ij} (second rating). Items with missing values are excluded from the calculation.

Mean Absolute Error (MAE)

$$\mathrm{MAE}_i = \frac{1}{10} \sum_{j=1}^{10} |a_{ij} - b_{ij}| \tag{1}$$

The range is 0–6, where values closer to 0 indicate higher consistency between the two ratings.

Exact Agreement Rate

$$\mathrm{Agree}_i = \frac{1}{10} \sum_{j=1}^{10} 1(a_{ij} = b_{ij}) \tag{2}$$

The range is 0–1, where 1 indicates perfect agreement across all items.

Furthermore, the mean values of MAE and agreement across the two duplicated words were calculated. Because human decision-making inherently involves uncertainty, it cannot be assumed that the two ratings for duplicated words would consistently

coincide. This uncertainty is considered to stem primarily from two sources. First, due to limitations in the information and computational resources available to individuals, "bounded rationality" emerges (Simon, 1955, 1972). In other words, individuals are unable to analyze situations in a fully rational and exhaustive manner, and instead rely on approximate judgments within the constraints of their cognitive resources. Second, cognitive processes such as perception, memory, and reasoning are inevitably subject to "information-processing noise" (Faisal et al., 2008; Kahneman et al., 2021), which introduces variability into judgments even under identical conditions. Through the interaction of these two factors, human decision-making is intrinsically variable. The thresholds were then established as follows.

- **High reliability:** $\mathrm{MAE} \leq 1.0$ (average deviation within one scale point) and $\mathrm{Agree} \geq 0.70$ (perfect agreement in more than 70% of cases).
- **Review required (gray zone):** $1.0 < \mathrm{MAE} \leq 2.0$ or $0.20 \leq \mathrm{Agree} < 0.70$.
- **Low reliability:** Otherwise.

3. Data Structure and Statistical Processing

3.1. Data Structure

Valid responses were obtained from 536 participants across 15 countries: China, South Korea, the Philippines, Myanmar, France, the United States, Vietnam, Malaysia, Thailand, Mongolia, Australia, Turkey, the United Kingdom, India, and Bangladesh. Of these, 273 responses (51%) were classified as high reliability, while 263 responses (49%) were classified as requiring further review.

Figure 2 illustrates the distributions of MAE and agreement, calculated from the two ratings for duplicated words and their averages. The histograms depict the frequency distributions of the two indices, while the overlaid kernel density estimation (KDE) curves highlight their distributional tendencies and modal peaks. The MAE exhibited a right-skewed distribution, with most cases concentrated below 1.0, indicating relatively small discrepancies between responses for duplicated words. By contrast, agreement displayed a bimodal distribution with peaks around 0.6 and 0.8, with a pronounced concentration near 0.8, suggesting that response consistency was generally high.

Moreover, after excluding outliers, the mean response time was 727.77 seconds, which fell within the expected range identified in the pilot study (600–1080 seconds). These findings indicate that

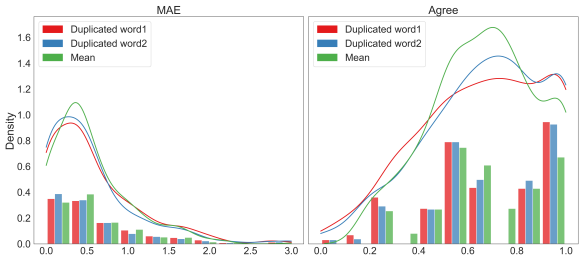

Figure 2: Distributions of MAE and agreement for duplicated words and their averages.

Learning level	Frequency	Rate
Intermediate	214	0.40
Upper Intermediate	147	0.27
Advanced	139	0.26
Upper Beginner	33	0.06
Beginner	3	0.01

Table 2: Distribution of participants' proficiency levels

the survey was conducted in a formally appropriate manner, with most participants devoting a reasonable amount of time to their responses. These results provide strong evidence supporting the overall reliability of the survey data.

Participants' ages ranged from 18 to 42 years, with a mean of 25 years and a standard deviation of 4.28. Regarding gender, 329 participants were female (61.38%), 205 were male (38.24%), and two identified otherwise. Self-assessed Japanese proficiency levels are summarized in Table 2. The largest group was Intermediate, comprising approximately 40% of the sample (214 participants), followed by Upper Intermediate at 27% (147 participants) and Advanced at 26% (139 participants). Overall, the majority of respondents (93%) fell between the intermediate and advanced levels, suggesting that most participants had already attained a relatively high level of Japanese proficiency.

3.2. Statistical Processing

Each valid response included 14 words, yielding a total of 7,504 word instances. However, because some words were rated by multiple responses, the final dataset contained 1,229 unique words. To enhance generalizability, it was desirable to collect as many responses as possible for each word. Therefore, in the present study, only words with at least five responses (N = 556) were retained for dataset evaluation.

The collected data were subject to potential biases stemming from both word-specific characteristics (e.g., abstractness) and respondent-specific tendencies (e.g., a general inclination to assign higher ratings). To address these sources of varia-

tion, ratings were estimated using a Bayesian Linear Mixed Model (BLMM), in which word-specific effects and participant-specific effects were incorporated as random effects.

The estimation was performed across ten dimensions: five related to familiarity (writing, reading, speaking, listening, and overall impression) and five related to meaning and image (negativity, calmness, imageability, abstractness, and understandability). All ratings were originally collected on a seven-point Likert scale (*Not at all* – *Strongly think so*) and were treated as continuous values ranging from 0 to 6.

Let $y_{ij}^{(k)}$ denote the rating given by participant j to word i on dimension k. The model was formalized as follows:

$$y_{ij}^{(k)} \sim \text{StudentT}(\nu, \mu_{ij}, \sigma), \tag{3}$$

$$\mu_{ij} = \alpha + \gamma_{\text{word}}^{(i)} + \gamma_{\text{subj}}^{(j)}. \tag{4}$$

$$\gamma_{\text{word}}^{(i)} \sim \mathcal{N}(\mu_{\text{word}}, \sigma_{\text{word}}), \quad \gamma_{\text{subj}}^{(j)} \sim \mathcal{N}(\mu_{\text{subj}}, \sigma_{\text{subj}}). \tag{5}$$

Here, α represents the overall mean, $\gamma_{\text{word}}^{(i)}$ denotes the word-specific effect, and $\gamma_{\text{subj}}^{(j)}$ denotes the participant-specific effect. These random effects were assumed to follow hierarchical normal distributions. In addition, a Student-t distribution was employed for the observation noise rather than a normal distribution, allowing for more robust estimation in the presence of outliers.

Parameter estimation was carried out using the No-U-Turn Sampler (NUTS). For each chain, 2,000 samples were drawn from the posterior distribution, with 1,000 iterations used for warm-up (tuning). Four independent chains were run in parallel, yielding a total of 8,000 samples. The target acceptance rate was set to 0.9 to suppress divergences and ensure stable convergence. Furthermore, the response variables were standardized (mean = 0, standard deviation = 1) to improve the stability of model estimation. As a result, the Gelman–Rubin convergence diagnostic statistic $\hat{R}$ was 1.0 for all parameters, confirming that all models successfully converged.

The estimation results are shown in Figure 3. The magnitudes of variation across words and across participants were represented by the standard deviations σ_{word} and σ_{subj}, respectively. The results indicated that the posterior means of both σ_{word} and σ_{subj} deviated from zero. Moreover, although some variability was observed in the ratings of individual words, differences in response tendencies across participants contributed more substantially. In other words, in the present dataset, individual differences among participants exerted a stronger influence on the ratings than did the inherent characteristics of the words themselves.

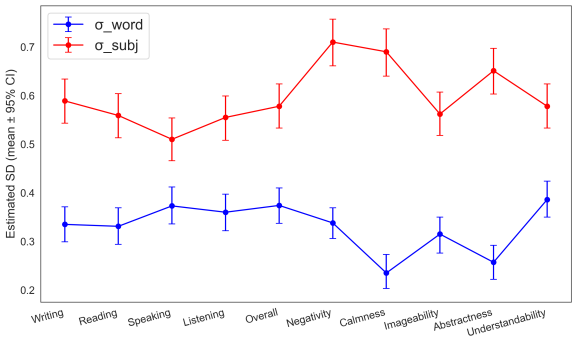

Figure 3: Estimated standard deviations of word and participant effects with posterior means and 95% credible intervals.

4. Evaluation and Benchmarking

Although the dataset is still under construction, this section sketches potential evaluation frameworks and benchmark experiments. Figure 4 illustrates the distribution of predicted mean scores across ten dimensions (Writing, Reading, Speaking, Listening, Overall, Negativity, Calmness, Imageability, Abstractness, and Understandability), with max-min normalization applied to constrain the data to the range of -3 and 3. Figure 5 presents a heatmap of the correlation matrix among the predicted mean scores for these dimensions. Based on the correlation structure in Figure 5, four distinct clusters were identified.

Interrelation of language skills The four language skills—writing, reading, speaking, and listening—showed very strong intercorrelations (0.81–0.90). Each skill was also highly correlated with Overall (0.89–0.94). This findings are consistent with previous research in second language acquisition and language assessment, which have emphasized the interdependent development of the four skills (Quinn et al., 2015; Zheng, 2024). The findings further confirmed, both statistically and conceptually, that Overall functions as an integrated index of the four core language skills.

Semantic properties and affective valence Abstractness was negatively correlated with both Imageability (–0.28) and Understandability (–0.17), while showing a positive correlation with Negativity (+0.22). This pattern suggests that abstract, less comprehensible, and less imageable words are more likely to be associated with negative affect.

Kousta et al. (2010) and Vigliocco et al. (2013) reported that abstract words are more strongly as-

sociated with affective dimensions such as valence and arousal than concrete words. This finding suggests that abstract words, which are difficult to visualize and require context-dependent interpretation, tend to evoke negative emotions in learners due to this dual processing burden. At the same time, it reflects the possibility that abstract words reinforce their meanings through the mediation of negative affect (Ponari et al., 2018). Taken together, these results support the view that affective valence is not merely ancillary but instead constitutes a critical factor that interacts with the abstractness—concreteness dimension in shaping cognitive processing. Thus, the observed correlation structure appropriately captures the intersection between semantic properties (abstract—concrete) and affective valence (negative—positive).

Understandability and imageability A strong positive correlation (0.76) was observed between Understandability and Imageability. This suggested that the semantic processing of words is reinforced through visual and sensory imagery, consistent with the psycholinguistic perspective that semantic representations are grounded in sensorimotor experience (Barsalou, 2008), a framework commonly referred as grounded cognition. Moreover, Understandability was highly correlated with each of the four language skills and with Overall (0.66–0.73), whereas Imageability showed moderate correlations with the skills and Overall (0.53–0.60). These results indicate that ease of processing—whether in terms of comprehension or imageability—is closely associated with language proficiency.

In summary, these results not only support established theories of lexical processing but also strengthen the validity of the present survey data.

Additionally, as a supplement to the correlations presented in Figure 5 and the identified four clusters, Figure 6 presents pairwise scatter plots among Overall, Negativity, Imageability, and Understandability. The scatter plots indicate that Overall proficiency is strongly predicted by both Imageability and Understandability, underscoring the central role of cognitive processing dimensions in lexical access. By contrast, Negativity exhibited only a weak negative correlation with Overall and no meaningful relationship with Imageability or Understandability, suggesting that affective valence constitutes a distinct psychological dimension. This pattern is consistent with psycholinguistic accounts in which abstract negative words rely on affective content to strengthen their semantic representation, while emotional load operates independently of cognitive processing efficiency. Taken together, these findings support the view that lexical processing can be conceived as a dual-route system, comprising a cognitive pathway and an affective pathway (Ferré et al., 2025; Kuperman et al., 2014; Palazova et al., 2013).

5. Conclusion

Japanese, often regarded as one of the most complex and difficult languages to acquire, remains insufficiently understood with respect to the factors that influence word difficulty and their relative contributions. Previous research has largely relied on corpus-based statistical measures, leaving a critical gap in data that directly capture the cognitive complexity underlying lexical processing. To address this limitation, the present study focused on six psycholinguistic dimensions—familiarity, affective, arousal, imageability, abstractness, and understandability—and developed a novel dataset through empirical surveys. Although data collection is still ongoing, preliminary quantitative evaluations have demonstrated the validity of the dataset, showing consistency with established theoretical frameworks and with findings from previous studies.

Word difficulty is not confined to surface-level lexical properties; it is also shaped by contextual factors, sociocultural influences, cognitive processing costs, and affective dimensions. Moreover, these factors do not operate independently but instead interact in complex ways to influence lexical processing. Many aspects of this process remain theoretically and empirically unexplained. Accordingly, the dataset constructed in this study not only contributes to word difficulty estimation but also provides a foundation for elucidating the interrelationships and interactions among its contributing factors.

Future research should address two important limitations of the present study. First, because the dataset is based on subjective metalinguistic ratings, the findings should be interpreted as reflecting learners' perceived lexical difficulty rather than objective processing difficulty. Further work should therefore examine how these metacognitive judgments relate to behavioral measures such as reaction time, recall, and eye-tracking. Second, the absence of a paired cross-linguistic dataset prevents direct evaluation of whether the observed psycholinguistic dimensions are language-universal or language-specific. Constructing comparable datasets across languages would help clarify which determinants of lexical difficulty generalize across linguistic contexts and which are shaped by language-specific properties.

Ethical Considerations

This study involved human participants who provided subjective ratings of related psycholinguistic

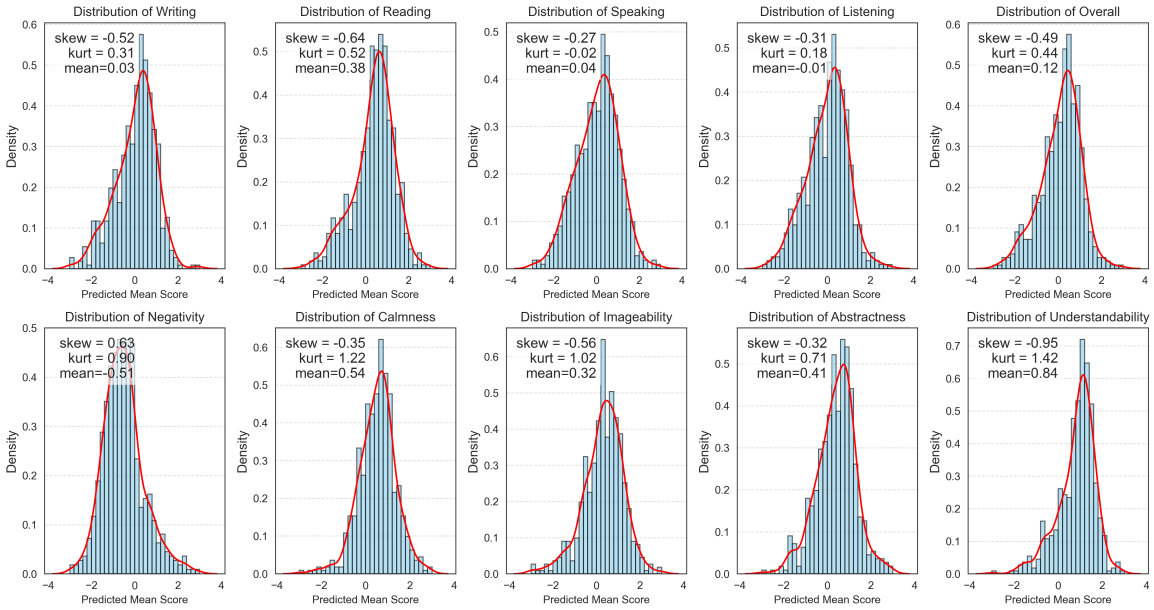

Figure 4: Distribution of predicted mean scores across ten dimensions (Writing, Reading, Speaking, Listening, Overall, Negativity, Calmness, Imageability, Abstractness, and Understandability). Each histogram depicts the density of predicted scores with an overlaid kernel density estimate (red line). The skewness (skew), kurtosis (kurt), and mean values are reported for each dimension.

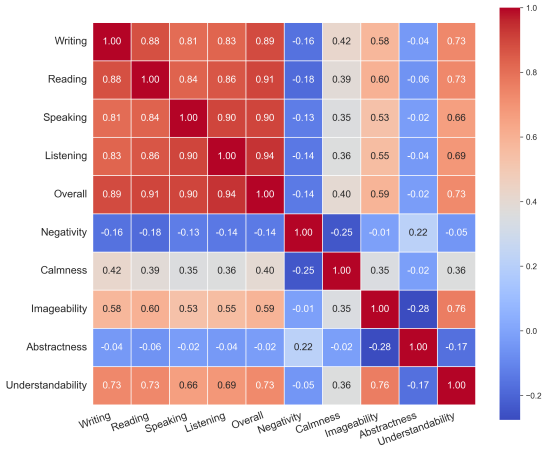

Figure 5: Correlation matrix of predicted mean scores across ten dimensions (Writing, Reading, Speaking, Listening, Overall, Negativity, Calmness, Imageability, Abstractness, and Understandability).

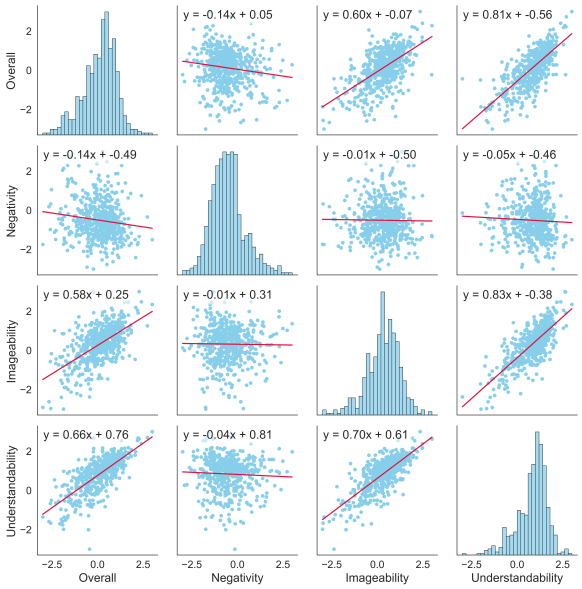

Figure 6: Pairwise scatter plots with regression lines among four dimensions (Overall, Negativity, Imageability, and Understandability).

properties. All participants gave informed consent prior to participation. The study protocol was reviewed and approved by the institutional ethics committee of Nagoya University.

While the dataset may be made publicly available for research purposes, all shared data will be fully anonymized and distributed under terms that restrict re-identification and misuse.

Acknowledgements

This work is supported by JSPS KAKENHI Grant Number 25K16328 and JSPS Topic-Setting Program to Advance Cutting-Edge Humanities and Social Sciences Research Grant Number JPJS00122674991.

Bibliographical References

L. F. Barrett. 2017. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, Boston, US.

Lawrence Barsalou. 2008. Grounded cognition. *Annual review of psychology*, 59:617–45.

J. D. Bransford, A. L. Brown, and R. R. Cocking. 2000. *How people learn: Brain, mind, experience, and school*. National Academy Press, Washington D.C., U.S.

Madeleine Bruce and Martha Ann Bell. 2022. Vocabulary and executive functioning: A scoping review of the unidirectional and bidirectional associations across early childhood. *Human Development*, 66:167–187.

Derek N. Canning, Stuart McLean, and Joseph P. Vitta. 2024. Relative complexity in a model of word difficulty: The role of loanwords in vocabulary size tests. *Studies in Second Language Learning and Teaching*, 14:631–659.

Gina N. Cervetti, Elfrieda H. Hiebert, P. David Pearson, and Nicola A. McClung. 2015. Factors that influence the difficulty of science words. *Journal of Literacy Research*, 47:153–185.

J.-M. Dewaele. 2015. From obscure echo to language of the heart: Multilinguals' language choices for (emotional) inner speech. *Journal of Pragmatics*, 87:1–17.

Fiona J. Duff, Gurpreet Reen, Kim Plunkett, and Kate Nation. 2015. Do infant vocabulary skills predict school-age language and literacy outcomes? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 56:848–856.

A. Aldo Faisal, Luc P. J. Selen, and Daniel M. Wolpert. 2008. Noise in the nervous system. *Nature Reviews Neuroscience*, 9(4):292–303.

Pilar Ferré, Alberto J., Sánchez-Carmona, Juan Haro, Rocío Calvillo-Torres, Jacobo Albert, and José Antonio Hinojosa. 2025. How does emotional content influence visual word recognition? a meta-analysis of valence effects. *Psychon Bull Rev*, 32:570–587.

Hung Tan Ha, Duyen Thi Bich Nguyen, and Tim Stoeckel. 2024. What is the best predictor of word difficulty? a case of data mining using random forest. *Language Testing*, 41:828–844.

Brett J. Hashimoto and Jesse Egbert. 2019. More than frequency? exploring predictors of word difficulty for second language learners. *Language Learning*, 69:839–872.

Elfrieda H. Hiebert, Judith A. Scott, Ruben Castaneda, and Alexandra Spichtig. 2019. An analysis of the features of words that influence vocabulary difficulty. *Education Sciences*, 9.

Regina Jucks and Elisabeth Paus. 2012. What makes a word difficult? insights into the mental representation of technical terms. *Metacognition and Learning*, 7:91–111.

Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Little, Brown Spark, New York.

Yu Kanazawa. 2021. *Replication and extension of an empirical study about foreign language formulaic familiarity database: Reliability and Emotionality*, Bulletin suisse de linguistique appliquée, pages 149–164.

Yu Kanazawa. 2023. Lexical and contextual emotional valence in foreign language vocabulary retention. *The Mental Lexicon*, 18:339–365.

Callula Killingly, Linda J. Graham, Haley Tancredi, and Pamela Snow. 2025. Reciprocal relationships among reading and vocabulary over time: a longitudinal study from grade 1 to 5. *Reading and Writing*, 38:605–625.

Walter Kintsch. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK.

Stavroula-Thaleia Kousta, Gabriella Vigliocco, David Vinson, Mark Andrews, and Elena Campo. 2010. The representation of abstract words: Why emotion matters. *Journal of experimental psychology. General*, 140:14–34.

Alper Kumcu and Robin L. Thompson. 2020. Less imageable words lead to more looks to blank locations during memory retrieval. *Psychological Research*, 84:667–684.

Victor Kuperman, Zachary Estes, Marc Brysbaert, and Amy Warriner. 2014. Emotion and language: Valence and arousal affect word recognition. *Journal of experimental psychology. General*, 143.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (taales): version 2.0. *Behavior Research Methods*, 50:1030–1046.

Gondy Leroy and David Kauchak. 2014. The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association*, 21.

M. Marschark and C. Cornoldi. 1991. *Imagery and Verbal Memory*, Imagery and cognition, pages 133–182. Springer, New York, US.

Marina Palazova, Werner Sommer, and Annekathrin Schacht. 2013. Interplay of emotional valence and concreteness in word processing: An event-related potential study with verbs. *Brain and Language*, 125(3):264–271.

C.A. Perfetti. 1999. *Comprehending written language: A blueprint of the reader*, The neurocognition of language, pages 167–208. Oxford University Press, New York, NY.

M. Ponari, C. F. Norbury, and G. Vigliocco. 2018. Acquisition of abstract concepts is influenced by emotional valence. *Developmental science*, 21(2).

Jamie Quinn, Richard Wagner, Yaacov Petscher, and Danielle Lopez. 2015. Developmental relations between vocabulary knowledge and reading comprehension: A latent change score modeling study. *Child Development*, 86(1):159–175.

James Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Sara A. Schmitt, David J. Purpura, and James G. Elicker. 2019. Predictive links among vocabulary, mathematical language, and executive functioning in preschoolers. *Journal of Experimental Child Psychology*, 180:55–68.

Herbert A. Simon. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118.

Herbert A. Simon. 1972. Theories of bounded rationality. In Charles B. McGuire and Roy Radner, editors, *Decision and Organization*, pages 161–176. North-Holland.

Jeffrey Stewart, Joseph P. Vitta, Christopher Nicklin, Stuart McLean, Geoffrey G. Pinchbeck, and Brandon Kramer. 2022. The relationship between word difficulty and frequency: A response to hashimoto (2021). *Language Assessment Quarterly*, 19:90–101.

Y. Sunakawa, J. Lee, and M. Takahara. 2012. The construction of a database to support the compilation of japanese learners dictionaries. *Acta Linguistica Asiatica*, 2(2):97–115.

Deborah Talmi, Ulrich Schimmack, Theone Paterson, and Morris Moscovitch. 2007. The role of attention and relatedness in emotionally enhanced memory. *Emotion (Washington, D.C.)*, 7:89–102.

John Truscott. 2015. Consciousness in sla: A modular perspective. *Second Language Research*, 31(3):413–434.

Gabriella Vigliocco, Stavroula-Thaleia Kousta, Pasquale Rosa, David Vinson, Marco Tettamanti, Joseph Devlin, and Stefano Cappa. 2013. The neural representation of abstract words: The role of emotion. *Cerebral cortex (New York, N.Y. : 1991)*, 24.

Joseph P. Vitta, Christopher Nicklin, and Simon W. Albright. 2023. Academic word difficulty and multidimensional lexical sophistication: An english-for-academic-purposes-focused conceptual replication of hashimoto and egbert (2019). *Modern Language Journal*, 107:373–397.

Erina Yamakawa and Keishi Tajima. 2024. Estimating word difficulty using wikipedia (in japanese). In *Proceedings of International Workshop on Database Engineering and Information Management*.

Wanwan Zheng. 2024. Estimating word difficulty using stratified word familiarity. *Cogent Arts and Humanities*, 11.