

Subjective and Hierarchical Annotation of Text Difficulty in Japanese

Daisuke Maekawa¹ Kazumasa Omura² Sora Tarumoto²
Shotaro Ishihara^{*2} Tomoyuki Kajiwara^{1,3}

¹Ehime University ²Nikkei Inc. ³The University of Osaka

{maekawa@ai., kajiwarat}@cs.ehime-u.ac.jp

{kazumasa.omura, sora.tarumoto, shotaro.ishihara}@nex.nikkei.com

Abstract

This study presents a difficulty-annotated corpus of Japanese newspaper articles. The annotation in our corpus has two key features. First, it is hierarchical, with difficulty levels labeled at the word, sentence, and document levels, respectively. Second, it is subjective, with 99 readers from diverse backgrounds labeling the difficulty levels based on their own perceptions. Detailed analysis of our corpus revealed correlations in difficulty levels across linguistic units, such as between words and sentences, and sentences and documents. We also found that text difficulty varies depending on reader attributes such as age and topics of interest. Furthermore, we trained Transformer encoders on this corpus and developed a personalized difficulty estimation model that takes into account both linguistic and attribute information. Experimental results demonstrate that incorporating reader attributes improves the performance of difficulty estimation and that multi-task learning across language units is beneficial. Our difficulty annotations are publicly available.¹

1 Introduction

Automatic text difficulty estimation constitutes a core task in computational linguistics (Collins-Thompson, 2014), supporting the retrieval of texts appropriate for learners at different proficiency levels (Vajjala and Meurers, 2012) and text simplification (Alva-Manchego et al., 2020). Recent data-driven advancements have yielded independent corpora for distinct linguistic units. Widely utilized resources include OneStopEnglish (Vajjala and Lučić, 2018) and Newsela (Xu et al., 2015) at the document level, CEFR-SP (Arase et al., 2022) at the sentence level, alongside CompLex (Shardlow et al., 2020, 2021) and CWI (Paetzold and Specia, 2016; Yimam et al., 2018) at the word level.

^{*}This work was conducted while the author was with Nikkei Inc.

¹<https://github.com/Nikkei/SHiAnn>

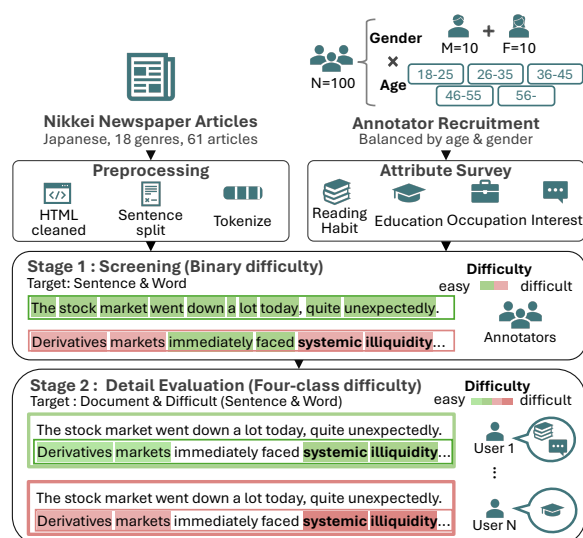


Figure 1: Overview of the annotation process. Although the actual data is in Japanese, it is presented in English for convenience.

However, existing literature predominantly treats the difficulty of these units in isolation, leaving the interrelationships across word, sentence, and document granularities under-explored. Existing resources typically focus on a single linguistic unit or adopt different annotation protocols across units, making cross-level analysis difficult. In particular, no corpus currently provides difficulty annotations across multiple linguistic units using the same annotators, the same four-level difficulty scale, and the same notion of subjective difficulty.

Furthermore, prior research frequently overlooks individual differences in difficulty perception. Difficulty is not dictated solely by objective linguistic features or uniform expert assessments (Vajjala and Lucic, 2019). Rather, perceived difficulty can vary across readers depending on their knowledge and experience. The growing recognition of reader-aware difficulty estimation is evidenced by explorations into personalized models for English (Lee and Yeung, 2018; Gooding and Tragut, 2022) and findings

Definition	Word	Sentence / Document
L0: Very Easy	Can use routinely	Can use naturally
L1: Easy	Understand the meaning	Understand fully, though not used naturally
L2: Difficult	Recognized, but meaning unknown	Contains a few incomprehensible parts
L3: Very Difficult	Unrecognized and meaning unknown	Contains multiple incomprehensible parts

Table 1: Definitions of subjective difficulty across the word, sentence, and document levels. These four-level criteria draw upon the concepts of receptive and productive knowledge from the Vocabulary Knowledge Scale (Paribakht and Wesche, 1993). Specifically, L1 reflects receptive understanding, whereas L0 additionally reflects productive familiarity sufficient for natural use. A four-class scale forces explicit difficulty assessments by eliminating neutral options.

that L1 backgrounds influence word difficulty perception among Japanese learners (Ide et al., 2023; Nohejl et al., 2024). Nevertheless, these efforts primarily target specific learner demographics or narrow sets of attributes, leaving a gap in the foundational resources required to quantitatively elucidate individual differences among diverse readers and develop comprehensive, personalized models.

To address these issues, this study constructs a hierarchically difficulty-annotated dataset in which the same annotators sequentially assess the subjective difficulty of words, sentences, and documents based on shared criteria. The dataset incorporates demographic attributes (e.g., age, gender, education, and occupation) alongside cognitive and behavioral characteristics (e.g., reading habits and topics of interest). This design facilitates a quantitative analysis of how readers’ background knowledge and linguistic hierarchy impact difficulty perception. Drawing on these analytical findings, our model integrates linguistic features with reader attributes, demonstrating the efficacy of personalized difficulty estimation over traditional one-size-fits-all approaches.

2 Corpus Construction

This study constructs a difficulty-annotated Japanese corpus incorporating annotator attributes and subjective difficulty assessments across hierarchical linguistic units. This resource facilitates attribute-aware difficulty estimation and hierarchical analysis across these units. The subsequent sections detail its design and statistical properties.

2.1 Annotation Design

Data Selection and Preprocessing The target texts are sourced from Nikkei newspaper articles,² selected for their minimal typographical noise, clear

hierarchical structure, and diverse topic coverage. To ensure sufficient textual volume and contextual information, 61 articles comprising at least two paragraphs were extracted. These articles span 18 genres selected based on the Nikkei Telecom article categorization scheme,³ ensuring broad lexical and contextual diversity. The full list of the 18 genres is provided in Appendix A. Preprocessing involved stripping HTML tags using BeautifulSoup⁴ to isolate the main body text. Sentence segmentation was performed using ja_sentence_segmenter,⁵ while word tokenization was performed using MeCab (Kudo et al., 2004) with the Neologd dictionary.⁶ To preserve semantic consistency at the word level, consecutive nouns were concatenated into single compound nouns.

Annotator Selection Annotators were recruited via the Lancers crowdsourcing platform.⁷ To ensure demographic diversity, 10 male and 10 female native Japanese speakers were targeted for each of the five age brackets (18–25, 26–35, 36–45, 46–55, and 56+), resulting in 99 final annotators who participated upon agreeing to the explained research purpose.⁸ The annotators received an hourly compensation of ¥1,200, surpassing the Japanese minimum wage. A supplementary questionnaire collected detailed demographic and behavioral attributes, including age and gender, as well as education, occupation, reading habits, and topics of interest. The

³Refer to Nikkei Telecom “Article Classification Keywords.” https://telecom.nikkei.co.jp/qanavi/reference/media/post_17

⁴<https://www.crummy.com/software/BeautifulSoup/>

⁵https://github.com/wwwcojp/ja_sentence_segmenter

⁶<https://github.com/neologd/mecab-ipadic-neologd>

⁷<https://lancers.jp/>

⁸Initially, 100 annotators were recruited; however, one male participant in the 18-25 age bracket withdrew, leaving 99 annotators.

²<https://nkbb.nikkei.co.jp/article/corpus/>

unit	total	L0 (%)	L1 (%)	L2 (%)	L3 (%)
Word	490,743	61.6	33.4	2.7	2.3
Sentence	86,031	39.1	50.8	8.7	1.4
Document	6,039	20.9	52.4	19.3	7.3

Table 2: Distribution of difficulty labels. The total count denotes the cumulative number of words, sentences, and documents flagged as “difficult” by the 99 annotators during Stage 1 and subsequently assigned four-class labels in Stage 2.

topics of interest were formulated as a multiple-choice selection from the 14 primary categories featured on the Nikkei Online Edition homepage.⁹ These attributes are subsequently used in the difficulty trend analysis and model development.

Two-Stage Annotation Words, sentences, and documents received four-class subjective difficulty assessments based on the criteria outlined in Table 1. Since exhaustive four-class evaluation across all words and sentences would impose a substantial annotation burden, a two-stage annotation pipeline was introduced to mitigate this burden and improve efficiency:

- **Stage 1 (Screening):** All words and sentences undergo a binary evaluation (easy/difficult). Any unit deemed “difficult” by at least one annotator is extracted for further assessment.
- **Stage 2 (Detailed Assessment):** The entire document, alongside the words and sentences evaluated as “difficult” in Stage 1, receives a granular four-class difficulty evaluation.

The annotation process utilized an annotation tool that displays the full document during word-, sentence-, and document-level evaluation. Details of the tool and the instructions provided to annotators are described in Appendix B. This design achieves both operational efficiency through target filtering and precise, context-aware labeling. We also recorded annotation times and manually inspected the collected annotations for obvious inconsistencies.

2.2 Statistics

Table 2 summarizes the corpus statistics. The dataset comprises 61 documents, 1,077 sentences, and 22,831 words, averaging 18 sentences per document and 21 words per sentence. The 99 anno-

tators universally evaluated all 61 articles, yielding a densely annotated dataset with approximately 490,000 word-level, 86,000 sentence-level, and 6,000 document-level difficulty labels. Reflecting the nature of newspaper articles, evaluations of “Very Easy” and “Easy” (L0, L1) account for the majority across all linguistic units. Conversely, the proportion of “Difficult” and “Very Difficult” (L2, L3) increases sequentially from words to sentences to documents, indicating that difficulty rises with the hierarchical level of the linguistic units.

The annotations are released under the CC BY-NC-SA 4.0 License.¹ The Nikkei newspaper articles used as the target texts are available for a fee. To protect privacy and prevent personal identification, all annotator data is strictly anonymized using unique IDs, and demographic attributes are limited to general categories.

3 Analysis

This section first evaluates annotation agreement to confirm the high subjectivity inherent in difficulty perception. Subsequently, it analyzes the impact of diverse reader attributes and the hierarchical relationships of linguistic units on perceived difficulty.

3.1 Agreement Analysis

Inter-annotator agreement, measured via Krippendorff’s α , yielded 0.177 at the word level, 0.076 at the sentence level, and 0.051 at the document level. Although these values are low, such divergence aligns with existing literature (Ide et al., 2023); text difficulty perception inherently fluctuates based on individual reading proficiency and lexical knowledge. Furthermore, the highly skewed label distribution detailed in Section 2.2 inflates the probability of chance agreement, thereby artificially deflating the resulting α scores.

Comparing agreement across linguistic units reveals a progressive decline from words to sentences and ultimately to documents. The relatively higher agreement at the word level likely stems from a widely shared baseline of lexical recognition. In contrast, sentence- and document-level difficulty evaluation may involve a broader range of factors, including cohesion, background knowledge, and other discourse-level characteristics (Graesser et al., 2004). The involvement of these multiple dimensions may, in turn, contribute to greater variability in subjective difficulty perception across readers. Thus, the low agreement should not be in-

⁹<https://www.nikkei.com/>

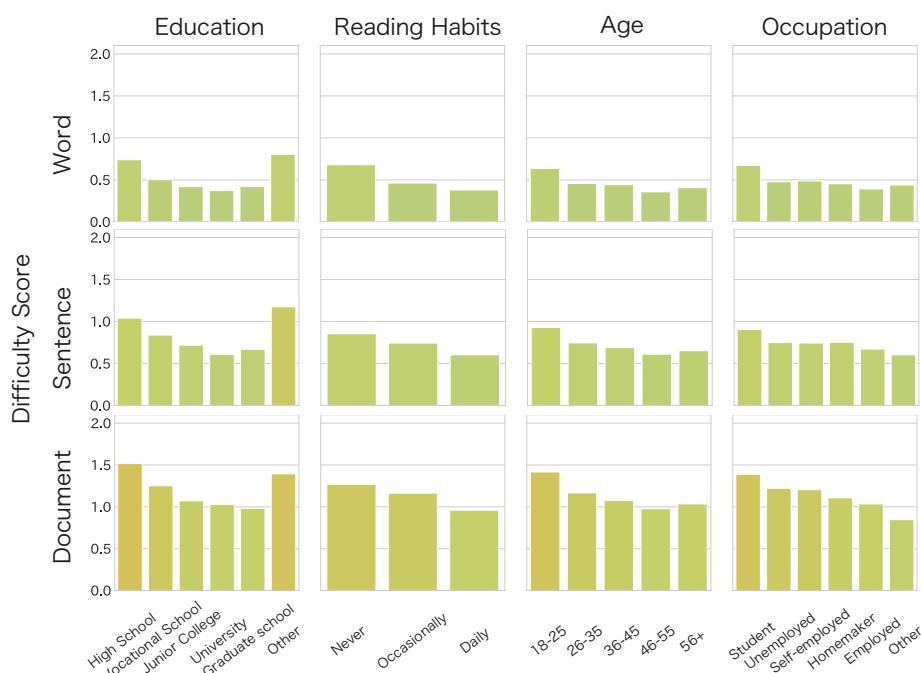


Figure 2: Average perceived difficulty by annotator attribute across linguistic units.

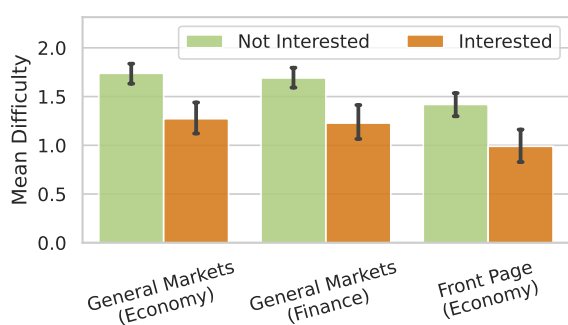


Figure 3: Differences in perceived difficulty based on reader interest. The x-axis represents the top three genre-interest pairs exhibiting the largest disparities in difficulty.

terpreted solely as annotation noise, but also as evidence of systematic differences in difficulty perception across readers. Consequently, these findings motivate modeling text difficulty as a reader-dependent perception rather than assuming a single gold-standard label.

3.2 Influence of Reader Attributes

This section analyzes age, education, occupation, reading habits, and topics of interest to elucidate their impact on difficulty perception. Figure 2 illustrates the distribution of average difficulty across linguistic units for each attribute.

Education and Reading Habits Attributes associated with linguistic proficiency and background knowledge exhibit a clear correlation with perceived difficulty. Regarding education, perceived difficulty consistently decreases as the education level rises. Similarly, habitual readers report lower difficulty levels compared to non-habitual readers. These findings substantiate that a reader’s vocabulary and background knowledge directly facilitate text comprehension.

Age and Occupation An analysis by age reveals that younger readers report the highest perceived difficulty, with scores gradually declining as age increases. This trend aligns with occupational differences, as students, who predominantly comprise the younger demographic, report significantly higher difficulty scores than working professionals in the middle-aged and older cohorts. This disparity likely stems from a relative lack of real-world experience and exposure to domain-specific knowledge among younger readers, rendering the corpus texts more challenging for them to comprehend.

Difficulty Trends Across Linguistic Units Notably, the aforementioned correlations between attributes and difficulty remain consistent across all linguistic units: words, sentences, and documents. Furthermore, while maintaining this consistency, the difficulty scores for each attribute progressively

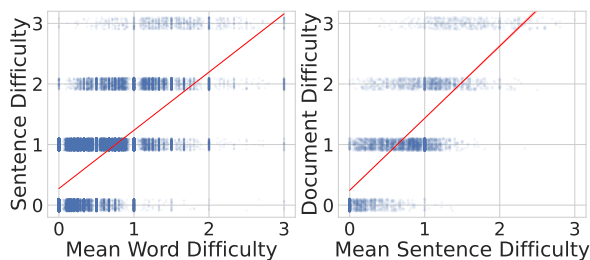


Figure 4: Scatter plots of perceived difficulty per annotator. Left: Sentence difficulty versus average word difficulty. Right: Document difficulty versus average sentence difficulty.

increase from words to sentences to documents. This trend inversely mirrors the agreement rate progression discussed in Section 3.1, where agreement declines as the linguistic unit expands. Together, these findings suggest that larger linguistic units impose higher cognitive processing loads, resulting in elevated difficulty and amplified individual variance.

Topics of Interest A subsequent analysis examines how an annotator’s topics of interest affect perceived document difficulty across different article genres. Figure 3 highlights the top three genres exhibiting the most substantial divergence in perceived difficulty based on the presence or absence of reader interest.

As shown in Figure 3, across all three highlighted genres, annotators possessing relevant interests rated the texts as significantly less difficult than those without such interests ($p < 0.001$). The most pronounced differences emerged in genres such as General Market and Front Page, which demand extensive and advanced knowledge in domains including finance, economics, and politics. Although these genres frequently feature specialized terminology and complex contexts, the prerequisite knowledge accompanying reader interest may effectively mitigate the cognitive load during text comprehension.

Conversely, general-interest articles such as Sports and Education yielded lower overall difficulty scores, with interest-based differences remaining marginal. This implies that highly accessible texts do not necessitate extensive prerequisite knowledge. Consequently, the comprehension benefits conferred by reader interest appear inherently contingent upon the absolute difficulty of the target text.

3.3 Difficulty in Linguistic Hierarchy

This section analyzes how difficulty propagates and aggregates across the distinct hierarchical levels of words, sentences, and documents.

To visually capture these interrelationships, Figure 4 illustrates the correlations between the average word difficulty and sentence difficulty, as well as between the average sentence difficulty and document difficulty, for each annotator. As depicted, the scatter plots reveal a pervasive positive correlation across the linguistic units. Specifically, a strong positive correlation exists between sentences and words ($r = 0.68$), alongside a similarly strong correlation between documents and sentences ($r = 0.73$). These high correlation coefficients substantiate that text difficulty does not exist in isolation; rather, it exhibits a hierarchical dependency following the progression from words to sentences to documents. In other words, the corpus demonstrates a hierarchical relationship wherein the presence of difficult words elevates sentence complexity, and the accumulation of difficult sentences subsequently amplifies overall document difficulty.

Conversely, instances exist where the overall sentence difficulty is rated high despite the low average difficulty of its constituent words. For instance, in the sentence “Based on the buying balance in margin trading (Tokyo, Osaka, and Nagoya three markets, combined institutional and general margin) and the financing amount, QUICK calculated the estimated profit/loss rate.”, the average word difficulty is 1.0 (Easy), whereas the sentence difficulty is rated substantially higher at 1.98 (Difficult). This disparity likely stems not merely from specific technical terms, but from complex syntactic structures, such as parenthetical insertions, that inflate the cognitive processing load. This asymmetrical divergence, where vocabulary is accessible, yet syntax and context remain challenging, clearly indicates that the difficulty of an entire sentence cannot be directly reduced to the sum of its constituent word difficulties.

3.4 Section Summary

The multifaceted analyses in this section yield two critical insights into difficulty perception. First, corroborating prior research (Gooding and Tragut, 2022), the quantitative findings confirm that difficulty is a relative metric heavily contingent upon a reader’s background and individual attributes. Second, while an overarching correlation governs the

propagation of difficulty from words to sentences to documents, hierarchical asymmetries, such as accessible vocabulary coupled with complex syntax and context, were also observed.

These insights expose the limitations of traditional models that predict uniform difficulty solely from textual features. Consequently, they highlight the necessity of incorporating both linguistic hierarchies and reader attributes. Drawing on these findings, the subsequent section develops a personalized difficulty estimation model.

4 Experiments

This section evaluates difficulty estimation models that incorporate reader attributes at the word, sentence, and document levels. Furthermore, it investigates whether multi-task learning, which leverages the correlations across linguistic units identified in Section 3.3, further enhances estimation performance. Finally, the individual contributions of each annotator attribute are analyzed to clarify their specific impact on the estimations.

4.1 Experimental Setup

Task Definition and Dataset The difficulty estimation task for each linguistic unit is formulated as a regression problem predicting continuous values within the $[0, 1]$ interval. Gold-standard labels, originally comprising four classes (0–3), were normalized to this $[0, 1]$ range, and Mean Squared Error (MSE) was employed as the loss function. To capture hierarchical difficulty dependencies, the multi-task learning setup pairs two adjacent linguistic units. At each optimization step, a mini-batch is sampled for each task, and the model is updated by minimizing the sum of the MSE losses from the two tasks. As all tasks share the same normalized difficulty scale of $[0, 1]$ and the same regression objective, their loss magnitudes are directly comparable. Therefore, no additional loss weighting is applied. The dataset was partitioned into training, validation, and test sets at the document level in an 8:1:1 ratio.

Model Architecture The base architecture employs ModernBERT-Ja-310M,¹⁰ a pre-trained Japanese language model. Input formats vary by linguistic unit: [CLS] context sentence [SEP] target word [SEP] for words, [CLS] target sentence [SEP] for sentences, and [CLS] entire article

[SEP] for documents. For difficulty estimation, the final hidden state vector of the [CLS] token serves as the primary feature representation. To integrate annotator attributes, the experiments compare the following configurations:

- **User Mean (mean):** Ignores textual features, directly utilizing the average difficulty score from each annotator’s training data as the predicted value for the test data.
- **Text Only (w/o):** Excludes attribute information, feeding only the [CLS] token vector directly into the regression layer.
- **Attribute Concatenation (w/):** Represents annotator attributes as one-hot vectors and concatenates them with the [CLS] token vector prior to the regression layer.
- **MLP-encoded Attributes (w/MLP):** Applies a non-linear transformation to the one-hot attribute vectors via a Multi-Layer Perceptron (comprising a linear layer and a ReLU activation function), concatenating the resulting 32-dimensional representation with the [CLS] token vector.
- **User ID (userID):** Appends a token representing the annotator’s unique ID to the input text sequence.

Notably, the multi-task learning models share encoder layers across tasks, while separate regression heads are used for each linguistic unit. In the w/MLP setting, the MLP layer used to transform annotator attributes is also shared across tasks. We evaluate two attribute integration configurations for multi-task learning: **w/** and **w/MLP**.

Training Parameters Optimization utilized AdamW (Loshchilov and Hutter, 2019) with batch sizes of 64 for word and sentence levels, and 16 for the document level. Learning rates were determined via a grid search ($\{2, 3, 5\} \times 10^{-5}$) on the validation set, selecting the value that minimized the RMSE for each model and linguistic unit configuration (refer to Appendix C for specific learning rates). These hyperparameter settings were applied to both the single-task and multi-task learning models. Training incorporated early stopping, terminating when the validation MSE loss failed to improve for three consecutive epochs.

¹⁰<https://huggingface.co/sbintuitions/modernbert-ja-310m>

unit	model	RMSE↓	r ↑	ρ ↑
Word	mean	0.198	0.434	0.471
	w/o	0.200	0.405	0.402
	w/	0.191	0.492	0.502
	w/MLP	0.178	0.585	0.617
	userID	0.167	0.654	0.663
	userID + w/	0.163	0.669	0.669
	userID + w/MLP	0.162	0.676	0.673
	w/ (W+S)	0.189	0.504	0.513
	w/MLP (W+S)	0.179	0.582	0.614
Sentence	mean	0.188	0.532	0.548
	w/o	0.211	0.294	0.264
	w/	0.199	0.433	0.426
	w/MLP	0.180	0.580	0.591
	userID	0.174	0.616	0.624
	userID + w/	0.174	0.614	0.621
	userID + w/MLP	0.174	0.622	0.627
	w/ (W+S)	0.199	0.439	0.428
	w/MLP (W+S)	0.179	0.582	0.596
Document	mean	0.232	0.519	0.506
	w/o	0.253	0.344	0.327
	w/	0.245	0.440	0.451
	w/MLP	0.246	0.408	0.400
	userID	0.215	0.582	0.570
	userID + w/	0.215	0.586	0.578
	userID + w/MLP	0.217	0.576	0.560
	w/ (S+D)	0.242	0.421	0.421
	w/MLP (S+D)	0.215	0.584	0.589

Table 3: Difficulty estimation results across linguistic units. For the model configurations, ‘mean’ denotes the average difficulty per annotator derived from the training set; ‘w/o’ relies solely on linguistic features; ‘w/’ and ‘w/MLP’ integrate annotator attributes; and ‘userID’ incorporates user-specific IDs. Parenthetical suffixes (W+S, S+D) appended to the model names indicate multi-task learning configurations across Word (W), Sentence (S), and Document (D) levels.

Evaluation Metrics Predictive performance is evaluated using Root Mean Square Error (RMSE) to measure absolute prediction error. Additionally, to assess the validity of relative rankings and relationships among texts, which is a critical aspect of subjective difficulty perception, the evaluation incorporates Pearson’s correlation coefficient (r) and Spearman’s rank correlation coefficient (ρ).

4.2 Results

This section evaluates the difficulty estimation models across the word, sentence, and document levels, validating baseline performance, the efficacy of annotator attribute integration, and the hierarchical cross-unit impact introduced by multi-task learning. Table 3 summarizes the evaluation results. Given the highly imbalanced label distribution, we additionally report label-wise RMSE for both single-task and multi-task models in Appendix D.

Single-Task Learning Experimental results indicate that the User Mean baseline (mean), which ignores textual features, outperformed the text-only model (w/o) across all linguistic units. This finding demonstrates that individual reader bias exerts a more dominant influence than the uniform difficulty of the text itself, thereby substantiating the necessity of incorporating reader attributes.

Models integrating explicit attribute information successfully capture this individual variance, though their effectiveness varies by linguistic unit. While the simple attribute concatenation model (w/) surpasses the user mean baseline at the word level, the sentence level requires non-linear transformations (w/MLP) to achieve similar success, and at the document level, explicit attributes alone fail to exceed the baseline. This may partly reflect the limited number of unique documents available for training and the greater subjectivity associated with document-level difficulty perception. In contrast, models incorporating the annotator’s userID consistently outperformed the baseline across all units. Indeed, Principal Component Analysis on the userID token representations revealed a high correlation between the first principal component and each annotator’s average difficulty on the test set ($\rho > 0.93$ for words and sentences; $\rho > 0.81$ for documents), validating its advantage in capturing individual tendencies. Ultimately, integrating explicit attributes and this powerful latent representation achieved very strong performance, highlighting the importance of synergistically combining them.

Qualitative analysis illustrates how the w/MLP model produces reader-dependent predictions by incorporating explicit attributes. For example, for the term “Floating exchange rate system,” the baseline predicts a uniform score across readers, whereas the w/MLP model outputs 0.72 for a reader uninterested in finance and 0.33 for a highly interested one. This difference reflects how reader-specific prediction tendencies derived from the attribute representations are incorporated into the final difficulty scores. It should be noted, however, that this does not imply a term-specific interaction between the text and individual attributes. In contrast, performance degrades for polarizing words, such as general proper nouns, for which the perception of difficulty depends more on idiosyncratic life experiences than on overarching attributes. This highlights that while explicit attributes are generally effective, relying solely on them carries an inherent limitation in fully capturing reader bias.

unit	model	RMSE↓	$r \uparrow$	$\rho \uparrow$
Word	w/o	0.200	0.405	0.402
	+ Age	0.199	0.423	0.422
	+ Gender	0.201	0.398	0.400
	+ Education	0.195	0.457	0.449
	+ Occupation	0.200	0.409	0.409
	+ Reading Habit	0.199	0.421	0.418
	+ Interests	0.193	0.470	0.472
	w/	0.191	0.492	0.502
Sentence	w/o	0.211	0.294	0.264
	+ Age	0.209	0.314	0.288
	+ Gender	0.212	0.287	0.259
	+ Education	0.207	0.345	0.330
	+ Occupation	0.212	0.282	0.261
	+ Reading Habit	0.212	0.297	0.275
	+ Interests	0.206	0.359	0.346
	w/	0.199	0.433	0.426
Document	w/o	0.253	0.344	0.327
	+ Age	0.259	0.233	0.237
	+ Gender	0.256	0.351	0.313
	+ Education	0.248	0.407	0.414
	+ Occupation	0.252	0.318	0.311
	+ Reading Habit	0.268	0.162	0.234
	+ Interests	0.253	0.334	0.345
	w/	0.245	0.440	0.451

Table 4: Analysis of individual annotator attribute contributions. Each attribute is independently integrated into the baseline model (w/o), which relies solely on textual features. Bold values indicate notable improvements over the baseline.

Multi-Task Learning Furthermore, multi-task learning (W+S, S+D) revealed asymmetric interactions across linguistic units: bottom-up information propagation improves performance, whereas top-down influence can act as noise. Bottom-up efficacy is evident in document-level estimation, where joint training with sentences (S+D) enables the w/MLP model to surpass the user mean baseline, as finer-grained sentence labels provide a beneficial inductive bias. Conversely, top-down influence does not consistently improve performance and tends to degrade it, a pattern observed at both the word level (W+S) and the sentence level (S+D). This suggests that simply summing losses causes coarse, macro-level labels to over-smooth local variations in lower units. These results corroborate the “hierarchical asymmetry of difficulty” identified in Section 3.3. Ultimately, effectively leveraging bidirectional information necessitates architectures that explicitly capture hierarchical structures and differences in granularity, rather than relying on simple task parallelization.

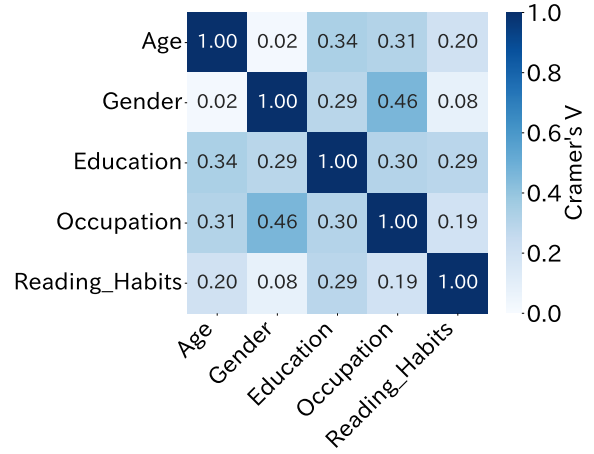


Figure 5: Heatmap illustrating the strength of association (Cramér’s V) among annotator attributes.

4.3 Attribute Analysis

To evaluate the individual contribution of each annotator attribute to subjective difficulty estimation, comparative experiments appended each attribute independently to the text-only baseline model, maintaining the configurations detailed in Section 4.1. Table 4 presents the evaluation results across the linguistic units.

As an overarching trend, the model integrating all annotator attributes (w/) achieves the highest predictive performance across the word, sentence, and document levels. Regarding individual attribute contributions, “education” and “topics of interest” consistently enhance the baseline performance across all linguistic units. This consistent improvement demonstrates that a reader’s objective knowledge proficiency and domain-specific interests play a pivotal role not only in local cognitive processing, such as resolving lexical or syntactic complexities, but also in macro-level document comprehension.

Conversely, the remaining attributes fail to improve performance when applied in isolation. Furthermore, even the effective attributes, namely education and interest, fall short of the comprehensive w/ model. These results reveal that the macro-level correlations between annotator attributes and average difficulty, as illustrated in Figure 2, do not necessarily translate into enhanced predictive accuracy for individual difficulty estimation. Ultimately, even if differences exist in average evaluation trends, a single attribute alone proves insufficient for modeling the intricate, highly individualized nature of difficulty perception.

The limitations of single-attribute prediction likely stem from confounding variables among the

attributes. To empirically quantify these interrelationships, Figure 5 visualizes the strength of association between categorical variables using Cramér’s V. The heatmap reveals moderate associations between specific pairs, such as “gender and occupation” and “age and education”, indicating that a reader’s background characteristics are deeply intertwined rather than strictly independent. Consequently, even if an attribute’s individual contribution appears marginal, comprehensively integrating multiple attributes remains important for highly accurate, personalized difficulty estimation.

5 Conclusion

This study constructed a difficulty-annotated corpus of Japanese newspaper articles spanning the word, sentence, and document levels, investigating the impact of annotator attributes and hierarchical cross-unit relationships on subjective difficulty perception. The findings demonstrate that, while explicit reader attributes alone have limitations in fully capturing individual bias, integrating multiple attributes is important for personalized difficulty estimation. Moreover, the strong performance of the w/MLP model suggests that non-linear transformation of attribute information can further improve prediction accuracy. Furthermore, multi-task learning experiments revealed asymmetrical dependencies across linguistic units; although bottom-up information propagation proves effective, fully exploiting these interactions necessitates hierarchical architectures that explicitly accommodate differences in information granularity.

Limitations

This study entails four primary limitations.

First, the current evaluation framework faces a cold-start problem regarding unseen users lacking historical reading data. Because this research primarily aims to elucidate how reader attributes and linguistic hierarchies impact difficulty perception to validate personalized estimation, the scope is deliberately restricted to existing users with sufficient historical data, the user group most likely to benefit in practical applications. Consequently, addressing unseen users falls outside the current scope. Nevertheless, the attribute-integrated model operates independently of reading history, thereby partially mitigating this issue for unseen users.

Second, the evaluations are confined to a specific domain and language, utilizing exclusively

a corpus of Japanese newspaper articles. Consequently, generalizability to other languages or informal texts, such as social media posts, remains unverified. However, the fundamental objective of this work lies in unraveling the underlying mechanisms of subjective difficulty perception, rather than claiming cross-domain or cross-lingual reproducibility. Thus, the specific choice of language and dataset does not constitute an intrinsic constraint on the core contributions.

Third, the corpus contains only 61 unique documents, which may limit generalization at the document level. This relatively small number of documents reflects our design choice to obtain dense personalized annotations by having all 99 participants annotate every document, yielding 6,039 document-level labels. Our primary objective was to construct a resource for analyzing subjective and personalized difficulty rather than to maximize the number of independent documents. Nevertheless, expanding the corpus to include a larger number of documents remains an important direction for future work.

Fourth, our annotation quality control has several limitations. Although we recorded annotation times and manually inspected the collected annotations for obvious inconsistencies, we did not employ explicit checks or repeated items to quantitatively assess annotator consistency. Future extensions of the corpus could incorporate such mechanisms to more directly assess annotation reliability.

References

- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. [Data-Driven Sentence Simplification: Survey and Benchmark](#). *Computational Linguistics*, 46(1):135–187.
- Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. [CEFR-Based Sentence Difficulty Annotation and Assessment](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219.
- Kevyn Collins-Thompson. 2014. [Computational Assessment of Text Readability: A Survey of Current and Future Research](#). *ITL - International Journal of Applied Linguistics*, 165:97–135.
- Sian Gooding and Manuel Tragut. 2022. [One Size Does Not Fit All: The Case for Personalised Word Complexity Models](#). In *Findings of the Association for Computational Linguistics*, pages 353–365.
- Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. [Coh-Metrix](#):

- [Analysis of Text on Cohesion and Language](#). *Behavior Research Methods, Instruments, & Computers*, 36:193–202.
- Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi, and Taro Watanabe. 2023. [Japanese Lexical Complexity for Non-Native Readers: A New Dataset](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 477–487.
- Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. [Applying Conditional Random Fields to Japanese Morphological Analysis](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.
- John Lee and Chak Yan Yeung. 2018. [Personalizing lexical simplification](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Adam Nohejl, Akio Hayakawa, Yusuke Ide, and Taro Watanabe. 2024. [Difficult for Whom? A Study of Japanese Lexical Complexity](#). In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability*, pages 69–81.
- Gustavo Paetzold and Lucia Specia. 2016. [SemEval 2016 Task 11: Complex Word Identification](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 560–569.
- T. Sima Paribakht and Marjorie Bingham Wesche. 1993. [Reading Comprehension and Second Language Development in a Comprehension-Based ESL Program](#). *TESL Canada Journal*, 11(1):09–29.
- Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. [CompLex — a New Corpus for Lexical Complexity Prediction From Likert Scale Data](#). In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with READING Difficulties*, pages 57–62.
- Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. [SemEval-2021 Task 1: Lexical Complexity Prediction](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation*, pages 1–16.
- Sowmya Vajjala and Ivana Lučić. 2018. [OneStopEnglish Corpus: A New Corpus for Automatic Readability Assessment and Text Simplification](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304.
- Sowmya Vajjala and Ivana Lucic. 2019. [On Understanding the Relation Between Expert Annotations of Text Readability and Target Reader Comprehension](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359.
- Sowmya Vajjala and Detmar Meurers. 2012. [On Improving the Accuracy of Readability Classification Using Insights From Second Language Acquisition](#). In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173.
- Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. [Problems in Current Text Simplification Research: New Data Can Help](#). *Transactions of the Association for Computational Linguistics*, 3:283–297.
- Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. [A Report on the Complex Word Identification Shared Task 2018](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78.

A News Article Genres

No.	Genre	No.	Genre
1	Business	10	General Markets
2	Career Development	11	International
3	Culture	12	Investment & Finance
4	Economics Classroom	13	Market Commodities
5	Education	14	Money & Investment
6	Front Page	15	Reading & Books
7	General	16	Society
8	General & Economy	17	Sports
9	General & Politics	18	Thematic Features

Table 5: The 18 news genres selected based on the Nikkei Telecom article categorization scheme.

B Annotation Details

This appendix provides additional details on the annotation tool and the instructions given to annotators.

B.1 Annotation Tool

Figure 6 shows the annotation tool developed in this study. To support context-aware difficulty evaluation, the interface continuously displays the full article at the top and provides a bottom panel for rating the target word, sentence, or document.

B.2 Annotation Instructions

We provided all annotators with the following instructions before annotation. The original instructions were written in Japanese; we provide an English translation below.

Task Overview. Please assess the difficulty of 61 newspaper articles based on your subjective judgment at three linguistic units: words, sentences, and articles.

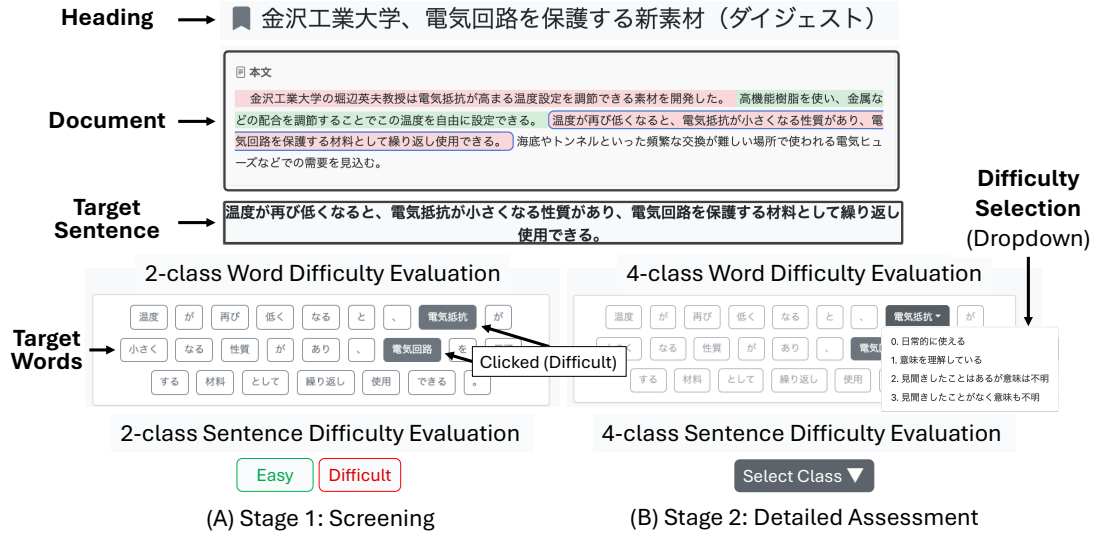


Figure 6: User interface of the annotation tool (A: Stage 1, B: Stage 2). The article shown on the screen is excerpted from *Nikkei newspaper articles*.

We aim to support the creation of newspaper articles that are easier to read for readers with diverse backgrounds and needs. We ask you to contribute to the construction of a dataset on the difficulty of newspaper articles.

Annotation Guidelines.

- Judge difficulty based on your first impression and intuition.
- Do not use search engines, dictionaries, or other external resources.
- A sentence may be judged as difficult even if it contains no difficult words.
- A sentence does not necessarily need to be judged as difficult even if it contains difficult words.
- If a word boundary appears unnatural, ignore the segmentation issue and judge the difficulty of the presented unit.
- In the four-level difficulty annotation, an unfamiliar word may still be judged as easy if you can understand its meaning.

C Hyperparameter Details

Table 6 presents the specific learning rates optimized on the validation set for each model and linguistic unit.

model	Word	Sentence	Document
w/o	2.0	2.0	3.0
w/	2.0	3.0	5.0
w/MLP	3.0	3.0	5.0
userID	5.0	2.0	2.0
userID + w/	2.0	3.0	3.0
userID + w/MLP	2.0	2.0	3.0
w/ (W+S)	2.0	2.0	-
w/MLP (W+S)	2.0	2.0	-
w/ (S+D)	-	5.0	5.0
w/MLP (S+D)	-	5.0	5.0

Table 6: Optimal learning rates ($\times 10^{-5}$) for each configuration.

D Label-wise Evaluation

Given the highly imbalanced difficulty-label distribution, Table 7 reports RMSE separately for each gold difficulty level. The results show that prediction errors tend to be larger for the minority high-difficulty labels across all linguistic units. Models incorporating annotator attributes or userID reduce errors for some high-difficulty labels. At the document level, multi-task learning with sentence-level difficulty further reduces RMSE for the high-difficulty labels. Overall, although minority high-difficulty labels remain challenging, incorporating reader information and multi-task learning can reduce prediction errors in some cases.

Table 7: Label-wise RMSE for each linguistic unit. The gold-label distributions (Very Easy / Easy / Difficult / Very Difficult) are 65/31/2/2% for words, 42/50/7/1% for sentences, and 26/54/15/5% for documents.

unit	model	Very Easy↓	Easy↓	Difficult↓	Very Difficult↓
Word	mean	0.147	0.146	0.453	0.773
	w/o	0.129	0.187	0.467	0.759
	w/	0.140	0.165	0.413	0.700
	w/MLP	0.124	0.147	0.413	0.689
	userID	0.126	0.136	0.363	0.608
	userID + w/	0.127	0.131	0.345	0.591
	userID + w/MLP	0.121	0.139	0.346	0.586
	w/ (W+S)	0.138	0.158	0.419	0.709
	w/MLP (W+S)	0.115	0.154	0.426	0.713
Sentence	mean	0.199	0.099	0.358	0.608
	w/o	0.206	0.125	0.419	0.718
	w/	0.186	0.126	0.399	0.680
	w/MLP	0.180	0.107	0.349	0.584
	userID	0.172	0.107	0.343	0.548
	userID + w/	0.172	0.109	0.342	0.545
	userID + w/MLP	0.179	0.108	0.325	0.521
	w/ (W+S)	0.210	0.113	0.365	0.637
	w/MLP (W+S)	0.174	0.112	0.353	0.586
	w/ (S+D)	0.195	0.122	0.388	0.661
	w/MLP (S+D)	0.155	0.130	0.375	0.602
Document	mean	0.305	0.140	0.234	0.451
	w/o	0.307	0.043	0.322	0.648
	w/	0.333	0.071	0.265	0.574
	w/MLP	0.333	0.094	0.269	0.544
	userID	0.258	0.120	0.248	0.472
	userID + w/	0.238	0.125	0.268	0.472
	userID + w/MLP	0.262	0.115	0.255	0.473
	w/ (S+D)	0.322	0.094	0.265	0.549
	w/MLP (S+D)	0.251	0.111	0.261	0.490