日本語テキスト難易度の主観的かつ多層的なアノテーション

前川 大輔 [1]　大村 和正 [2]　樽本 空宙 [2]　石原 祥太郎 [2]　梶原 智之 [1,3]
[1] 愛媛大学大学院理工学研究科　[2] 株式会社日本経済新聞社　[3] 大阪大学 D3 センター
{maekawa@ai., kajiwara@}cs.ehime-u.ac.jp
{kazumasa.omura, sora.tarumoto, shotaro.ishihara}@nex.nikkei.com

概要

本研究では，読み手の知識や能力に依存する主観的なテキストの難易度を詳細に分析するために，日本語の新聞記事における単語・文・文書の言語単位を対象とする難易度付きコーパスを構築する．従来の研究では，単語・文・文書の難易度は独立に扱われてきたが，本研究では同一のアノテータが共通の枠組みで各言語単位へ難易度を付与し，それらの間の関係を明らかにする．具体的には，複数の年齢層および性別のアノテータを採用し，言語単位ごとに 4 クラスの難易度を付与したコーパスを構築し，アノテータ属性と難易度の間の関係や言語単位間の難易度の相関を明らかにする．さらに，言語情報と属性情報の両方を考慮する難易度推定モデルを構築し，読み手に特化した難易度推定の有効性を示す．

1　はじめに

難易度推定は，テキストの難易度を計算論的に評価する計算言語学の主要な研究課題のひとつである [1]．先行研究では，単語・文・文書といった言語単位ごとに独立したコーパスが構築されており，文書レベルでは OneStopEnglish [2] や Newsela [3]，文レベルでは CEFR-SP [4]，単語レベルでは CompLex [5,6] や CWI [7,8] などがある．

これらの先行研究では，各言語単位の難易度が個別の対象として扱われており，単語・文・文書の異なる粒度間における難易度の関係性については，充分に検討されていない．難解な語彙の出現や文構造の複雑さは文書の可読性に影響を与えると考えられるが，これらを同一の枠組みで分析可能な言語資源は限られており，複数の言語単位を横断して一貫した基準で難易度を付与したコーパスは存在しない．

また，これらの先行研究では，難易度判断における読み手の主観的ばらつきや個人差が充分に考慮されていない．テキストの難易度は，読み手の知識や経験などに依存するため，同一のテキストであっても読み手によって感じられる難易度は異なる．例えば，日本語学習者における単語難易度の判断が，母語の違いによって大きく異なることが報告されており [9,10]，読み手の属性を考慮した難易度推定の重要性が示唆されている．

これらの課題に対処するために，本研究では，日本語の新聞記事を対象として，単語・文・文書の複数の言語単位に対する主観的な難易度アノテーションを，統一的な枠組みで付与したコーパスを構築する．先行研究とは異なり，同一のアノテータが共通の基準で各言語単位の難易度を判断することで，言語単位間における難易度の関係を一貫した形で分析可能とする．また，複数の年齢層および性別のアノテータを採用することで，アノテータ属性と難易度判断の関係を定量的に明らかにする．さらに，構築したコーパスを用いて，言語的特徴に加えて読み手の属性情報を考慮する難易度推定モデルを構築し，読み手に特化した難易度推定の有効性を検証する．

2　コーパス構築

本研究では，アノテータ属性を収集し，単語・文・文書の階層的な言語単位に対し，主観的な難易度を付与した日本語難易度付きコーパスを構築した．本コーパスは，アノテータ属性に基づく難易度推定や，言語単位間の階層的な分析を可能とする言語資源として設計されている．以下，その設計および統計情報について述べる．

2.1　アノテーション設計

データ選定および前処理　ソースデータには，「日本経済新聞記事オープンコーパス」[1)] を用い，十分なテキスト量と文脈情報を確保するため，2 段落以上で構成される 61 記事を抽出した．これらは，語

1）https://nkbb.nikkei.co.jp/article/corpus/

表 1　単語・文・文書に対する主観的難易度の定義．Paribakht and Wesche [11] の Vocabulary Knowledge Scale を参考に，受容知識と産出知識の区別に基づき 4 段階の基準を策定した．中間的な「普通」の選択を避け，難易度判断を明確に誘導するため，4 クラスの尺度を採用した．

定義	単語	文・文書
とても易しい	日常的に使える	自分でも自然に使える
易しい	意味を理解している	自分では使わないが意味は理解できる
難しい	見聞きしたことはあるが意味は不明	意味のわからない部分が少しある
とても難しい	見聞きしたことがなく意味も不明	意味のわからない部分が複数ある

彙や文脈の多様性を確保するため記事分類体系[2]に基づく全 18 ジャンルを網羅している．テキストの前処理として，BeautifulSoup[3] を用いて HTML タグを除去し，本文テキストのみを抽出した．抽出したテキストにおいて，文の抽出および単語分割には，日本語文分割器 ja_sentence_segmenter[4] と MeCab[5]（Neologd 辞書）をそれぞれ使用した．なお，単語単位の意味の一貫性を保つため，連続する名詞は複合名詞として 1 語に連結する処理を施した．

アノテータ選定　アノテータは，クラウドソーシングサービス Lancers[6] を通じて募集した．属性の多様性を確保するため，5 つの年齢区分（18-25 歳，26-35 歳，36-45 歳，46-55 歳，56 歳以上）から男女各 10 名ずつを選定し，計 99 名[7]を採用した．また，アノテータの属性を詳細に把握するため，年齢・性別に加え，学歴・職種・読書習慣・関心分野に関するアンケート調査を実施した．これらは後の難易度評価の分析とモデル構築のために用いる．

2 段階アノテーション　単語・文・文書に対し，表 1 の基準で 4 クラスの主観的難易度を付与した．全ての単語および文に対し網羅的に 4 クラスで難易度を評価することは，アノテータへの認知的負荷が高く，コスト面でも非効率であるため [12]，負荷軽減と効率化を可能とする 2 段階のアノテーション手法を導入した．

- **第 1 段階（スクリーニング）：**全単語・全文を対象に「易しい／難しい」の 2 クラスで評価する．アノテータが難解と感じた箇所を網羅的に抽出する．
- **第 2 段階（詳細評価）：**文書全体，および第 1 段階で「難しい」と判定された単語・文を対象に，4 クラスで詳細に難易度を評価をする．

これらの作業は，付録 A に示したアノテーションツールを用い，文書全体の文脈を提示しながら文や単語を評価するよう実施した．これにより，評価対象の絞り込みによる効率化と，文脈を考慮した精緻なラベル付与を両立させた．

2）日経テレコン「記事分類キーワード」を参照．https://t21help.nikkei.co.jp/reference/cat397/post-17.html
3）https://www.crummy.com/software/BeautifulSoup/
4）https://github.com/wwwcojp/ja_sentence_segmenter
5）https://github.com/neologd/mecab-ipadic-neologd
6）https://lancers.jp/
7）当初 100 名を採用したが，期間中に 18-25 歳男性枠の 1 名が離脱したため，最終的なアノテータ数は 99 名となった．

表 2　本コーパスの基本統計および難易度ラベルの分布．総数は 99 名による 4 クラスの難易度の付与数を示す．

言語単位	総数	L0 (%)	L1 (%)	L2 (%)	L3 (%)
単語	490,743	61.6	33.4	2.7	2.3
文	86,031	39.1	50.8	8.7	1.4
文書	6,039	20.9	52.4	19.3	7.3

2.2　統計情報

構築したコーパスの統計を表 2 に示す．本コーパスは 61 文書，1,077 文，22,831 単語で構成され，平均して文書あたり 18 文，文あたり 21 単語を含む．アノテーションの結果，単語・文・文書の各言語単位に対し，4 クラス難易度のラベルを収集した．難易度分布は，新聞記事というドメインの性質を反映し，いずれの単位も「易しい（L0, L1）」判定が過半数を占める．一方で，単語・文・文書の順に「難しい，とても難しい（L2, L3）」の割合が増加する傾向にあり，言語単位の階層化に伴う難易度の上昇が確認できる．

なお，日本経済新聞記事オープンコーパスは現時点で研究利用において無償で公開されており、本研究でのアノテーションも公開予定である．

3　分析

本節では，構築したコーパスに対して，言語単位間の関係性および難易度評価における主観性の影響について分析する．

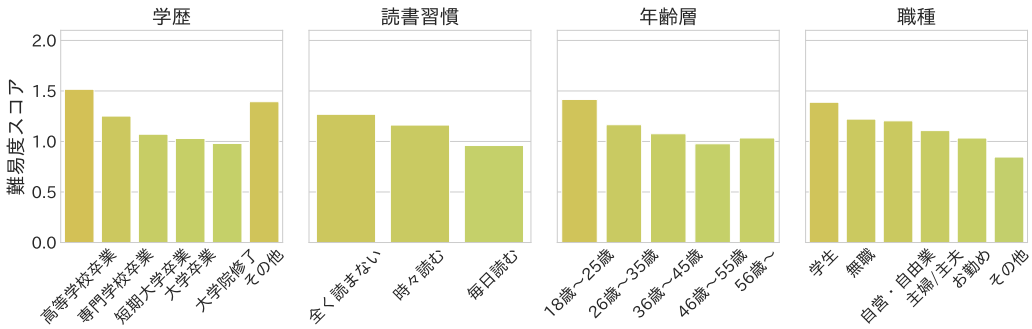


図 1　文書レベルにおける各属性ごとの平均難易度

3.1　一致率の分析

各言語単位における難易度評価の一致度を Krippendorff's α により評価したところ，単語レベルは 0.177，文レベルは 0.076，文書レベルは 0.051 であった．いずれも高い一致率ではないが，難易度評価が読解力や語彙知識などの個人差に強く依存するという既知の知見 [9] と整合する結果である．さらに，言語単位ごとの一致率を比較すると，**単語** > **文** > **文書** の順で一致率が低下していることがわかる．単語レベルでは，語彙の認知度という比較的共有しやすい基準が存在するため，ある程度の一致が見られたと考えられる．一方，文や文書レベルへと単位が大きくなるにつれて，単語同様に読者の背景知識に左右されることに加え [13]，文脈の複雑さや論理構成など，考慮すべき要素が多次元化するため，個人の主観による難易度評価の分散がより顕著になったと推察される．この分析結果は，単一の正解ラベルを予測する従来のモデルでは読み手に特化した高精度な推定が困難である可能性を示している．

3.2　属性情報と難易度の関係

アノテータ属性が難易度判定に与える影響を明らかにするため，年齢層，読書習慣，関心トピック，職種，および学歴，性別を対象に分析した．

まず言語単位間の関係に着目すると，属性ごとの相対的な難易度傾向は，単語・文・文書の全言語単位において高い一貫性が確認された．一方で，難易度の平均値は，その一貫性を保ったまま **単語** < **文** < **文書** の順で上昇する傾向が見られた．これは 3.1 節で述べた一致率の順序（単語 > 文 > 文書）と関係し，言語単位が大きくなるにつれて情報の処理負荷が増大し，より難易度が高く，かつ個人差が大きいことを示唆している．

以上の分析により，全単位で属性間に一貫した傾向が観察されたため，属性による差異が最も顕著に表れた文書レベルの結果を図 1 に示す．

学歴および読書習慣の影響　読み手の言語能力や背景知識に関連する属性は，難易度評価と相関を示した．学歴に関しては，教育水準が上がるにつれて，難易度評価が有意に低下する傾向が見られた．同様に読書習慣においても，習慣的に本を読む層の方が，そうでない層と比較して難易度を低く評価する傾向が確認された．これらの結果は，読み手の保有する語彙力や背景知識量が，テキストの理解に直接的な影響を与えていることを裏付けている．

年齢および職業による差異　年齢層別では，若年層の平均難易度が最も高く，年齢が上がるにつれて緩やかに難易度が低下する傾向が見られた．この結果は職業別の傾向とも整合しており，若年層に属する「学生」による難易度評価は，中年層から高年層を占める「有職者」と比較して顕著に高い値を示している．若年層や学生は，社会経験や特定ドメインへ触れる機会が相対的に低いため，コーパスに含まれるテキストをより難解に感じたと考えられる．

本分析により，難易度評価は読み手の属性に強く依存し，テキスト特徴のみに基づく従来手法では個人差を捉えきれないことが判明した．すなわち，高精度な推定には主観性の考慮が不可欠である．

3.3　各言語単位間の難易度の階層的関係

本節では，単語・文・文書という異なる階層の言語単位間において，難易度がどのように伝播・構成されているかを分析する．

各単位間の関係を定量化するため，単語・文・文書レベルの平均難易度間の相関分析を行った．Pearson の相関係数を算出した結果，文と単語の間には強い正の相関（$r = 0.68$）が見られ，文書と文の間にはさらに強い相関（$r = 0.74$）が確認された．この高い相関係数は，テキストの難易度が独立して

存在するのではなく，「**単語 → 文 → 文書**」という階層的な関係を持っていることを裏付けている．すなわち，難解な単語の連続が文を難化させ，難解な文の蓄積が文書全体の難易度を引き上げるという階層的な依存関係が，本コーパスにおいて定量的に実証され，言語単位を統合して難易度を付与したアプローチの有効性を示している．

4 評価実験

本研究では，属性情報が難易度に与える影響を検証するため，単語・文・文書の各言語単位における難易度推定性能を評価した．

4.1 実験設定

タスク定義とデータセット 本実験では，単語・文・文書の各言語単位に対する難易度推定を，[0, 1] 区間への回帰問題として定式化した．正解ラベルは 4 クラスのラベル (0-3) を [0, 1] 区間に正規化し，損失関数には MSE 損失を用いた．データセットは記事単位で訓練・検証・評価を 8 : 1 : 1 に分割した．

モデル構造 ベースモデルには，日本語事前学習済みモデルである ModernBERT [8] を採用した．モデルへの入力として，テキスト情報に加え，アノテータ属性（One-hot ベクトル）をテキスト埋め込みと結合してエンコーダに入力する設定（w/）と，テキストのみの設定（w/o）を比較した．各言語単位の入力形式は以下の通りである．

- **単語:** `[CLS]` 文 `[SEP]` 対象単語 `[SEP]`
- **文:** `[CLS]` 対象文
- **文書:** `[CLS]` 記事全体

学習パラメータ 最適化手法には AdamW [14] を用い，バッチサイズは単語および文レベルで 64，文書レベルで 16 とした．学習率は検証データを用いたグリッドサーチ（$\{2, 3, 5\} \times 10^{-5}$）により決定し，各設定において RMSE が最小となる値を採用した．具体的な学習率は以下の通りである．単語 (w/) と文書 (w/o) には 2.0×10^{-5}，単語 (w/o) には 3.0×10^{-5}，それ以外には，5.0×10^{-5} を適用した．学習は，検証データの MSE 損失が 3 エポック連続で改善しなかった場合に Early Stopping を適用した．

評価指標 モデルの予測性能を定量的に評価するため，RMSE，Pearson の相関係数，および Spearman の順位相関係数の 3 つの指標を採用した．

8） https://huggingface.co/sbintuitions/modernbert-ja-310m

表 3 各言語単位における難易度推定の結果．w/o が言語特徴のみの場合，w/が言語特徴とアノテータ属性を用いた場合を示す．

言語単位	モデル	RMSE↓	Pearson↑	Spearman↑
単語	w/o	0.227	0.428	0.411
	w/	**0.223**	**0.495**	**0.504**
文	w/o	0.213	0.257	0.235
	w/	**0.204**	**0.372**	**0.356**
文書	w/o	0.262	0.164	0.212
	w/	**0.253**	**0.346**	**0.348**

4.2 結果

単語・文・文書の各言語単位に対して難易度推定を行い，ベースライン性能およびアノテータ属性情報の有効性を検証した．表 3 に，各言語単位における評価結果を示す．

実験の結果，アノテータ属性を利用したモデル（w/）は，利用しないモデル（w/o）と比較して，全ての言語単位において RMSE，Pearson 相関係数，Spearman 順位相関係数の全指標で性能が向上した．この結果は，アノテータ属性の導入が，予測値の絶対的な誤差を最小化するだけでなく，難易度の相関においても，言語単位問わず一貫して有効であることを示している．特に文書のような長い単位においても，読み手の属性情報を考慮することで，主観的な難易度をより正確に捉えられることが明らかとなった．

5 おわりに

本研究では，日本語新聞記事に対して，単語・文・文書の各言語単位に主観的な難易度を付与したコーパスを構築し，アノテータ属性と言語単位間の階層関係が難易度に与える影響を分析した．実験の結果，読書習慣や年齢といった属性によって難易度に一定の傾向が見られ，属性情報が難易度推定に有効であることが示された．また，各言語単位間の難易度には強い相関が確認され，言語単位間の難易度情報を活用した各言語単位の難易度推定の可能性を示唆している．今後は，言語単位間の相関関係を考慮した階層的なモデルやマルチタスク学習を導入し，言語単位間の難易度情報が性能に寄与するのかを検証する．

参考文献

[1] Kevyn Collins-Thompson. Computational Assessment of Text Readability: A Survey of Current and Future Research. **ITL - International Journal of Applied Linguistics**, Vol. 165, pp. 97–135, 2014.

[2] Sowmya Vajjala and Ivana Lučić. 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**, pp. 297–304, 2018.

[3] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in Current Text Simplification Research: New Data Can Help. **Transactions of the Association for Computational Linguistics**, Vol. 3, pp. 283–297, 2015.

[4] Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. CEFR-Based Sentence Difficulty Annotation and Assessment. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 6206–6219, 2022.

[5] Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 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**, pp. 57–62, 2020.

[6] Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. SemEval-2021 Task 1: Lexical Complexity Prediction. In **Proceedings of the 15th International Workshop on Semantic Evaluation**, pp. 1–16, 2021.

[7] Gustavo Paetzold and Lucia Specia. SemEval 2016 Task 11: Complex Word Identification. In **Proceedings of the 10th International Workshop on Semantic Evaluation**, pp. 560–569, 2016.

[8] Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 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**, pp. 66–78, 2018.

[9] Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi, and Taro Watanabe. 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**, pp. 477–487, 2023.

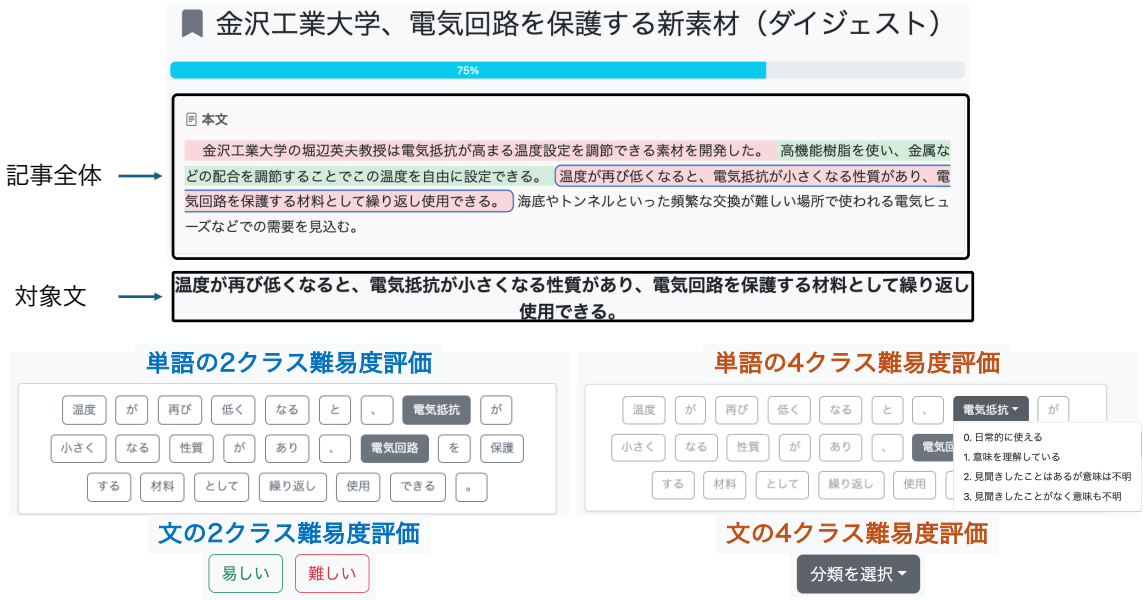
[10] Adam Nohejl, Akio Hayakawa, Yusuke Ide, and Taro Watanabe. Difficult for Whom? A Study of Japanese Lexical Complexity. In **Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability**, pp. 69–81, 2024.

[11] T. Sima Paribakht and Marjorie Bingham Wesche. Reading Comprehension and Second Language Development in a Comprehension-Based ESL Program. **TESL Canada Journal**, Vol. 11, No. 1, p. 09–29, 1993.

[12] Matthew Shardlow, Richard J. Evans, and Marcos Zampieri. Predicting Lexical Complexity in English Texts: The Complex 2.0 Dataset. **Language Resources and Evaluation**, Vol. 56, pp. 1153 – 1194, 2021.

[13] Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. Coh-Metrix: Analysis of Text on Cohesion and Language. **Behavior Research Methods, Instruments, & Computers**, Vol. 36, pp. 193–202, 2004.

[14] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In **Proceedings of the Seventh International Conference on Learning Representations**, 2019.



(A) 1段階目：スクリーニング　　(B) 2段階目：詳細評価

図 2　アノテーションツールの UI（A：1 段階目，B：2 段階目）．画面内の記事は『日本経済新聞記事オープンコーパス』から抜粋

A　アノテーションツールの詳細

本研究で構築したアノテーションツールのインターフェースを図 2 に示す．アノテータが文脈を考慮して難易度を評価できるよう，画面上部に記事全体を常時表示し，評価対象となる単語・文・文書を下部の操作パネルで難易度評価する UI を採用した．