日本語能力試験に基づく日本語文の難易度推定

中町 礼文 [1]　佐藤 敏紀 [1]　西内 沙恵 [2]　浅原 正幸 [2]　奥村 学 [3]
[1]LINE 株式会社　[2] 国立国語研究所　[3] 東京工業大学
{akifumi.nakamachi,toshinori.sato}@linecorp.com
{snishiuchi,masayu-a}@ninjal.ac.jp
oku@pi.titech.ac.jp

概要

急増する在留外国人に向けての日本語の学習支援や,「やさしい日本語」の整備という社会的課題がある. 特に, 在留外国人の増加に伴い, 日本語教師の負担が増大しており, 学習支援の自動化は急務である. 本研究では, 非母語話者の日本語運用能力の評価試験として広く用いられている日本語能力試験 (JLPT) の習熟度基準に準拠した日本語文の難易度推定の自動化に取り組む. まず, JLPT の過去の試験問題から評価用データを作成し, JLPT の学習参考書から訓練用データを作成した. さらに, 作成したデータをもとに, JLPT の習熟度基準に基づいた文の難易度の推定器を作成した. 我々の作成した推定器は評価データによる評価で高い相関を示した.

1　はじめに

在留外国人の 76% は, 日常でのコミュニケーションで日本語を用いている [1]. 一方で, 日本語教師の数は横ばい [2] であり, 計算機による日本語の学習・教育支援の自動化が急がれている. 日本語の難易度判定は, 非母語話者の理解力に合わせたテキスト提示を行う際のフィルターや, 作文練習の自動支援など, 日本語の教育支援に幅広く活用される. 例えば, コミュニケーションアプリケーション上で, 日本語の記述練習の支援を行うことや, ニュースなどの情報配信サービス上で文の難易度を提示することで, 在留外国人の日本語での日常生活に即した学習を支援できる.

テキストの難易度推定の既存研究として, 英語におけるリーダビリティの研究では, 単語の平均音節数や, 平均文長などをもとに文章の難易度を評価する FKGL[3] を始め, 多数の評価方法が開発されている. また, CEFR-J に基づく英語の文章の難易度を評価する手法 [4, 5] などもあるが, 英語と日本語は言語的性質が異なるため, 日本語の難易度評価は英語の難易度評価とは異なる点が多い. 日本語の難易度評価の既存研究としては, 平均文長や単語難易度などの基本的な統計量をもとに文章の難易度を推定する回帰式 [6, 7, 8, 9, 10] が多数存在している. 李 [9] は, BCCWJ から独自にコーパスを作成し, 平均文長や, 漢語率などの統計量から文章の難易度を推定する回帰式を作成し, ニュースなどの一般文書の難易度評価実験を行った. また, 国語教育を対象とした文章の難易度判定システムとして, 日本語の教科書を用いた大規模なコーパスの n-gram や統計的言語モデルに基づく特徴量を利用して, 初等教育の学年を推定する手法 [11, 12] がある. また, 日本語の語彙の難易度判定システムとして, Sunakawa ら [13] は, 人手で単語難易度の辞書を作成した. Nishihara ら [14] は, 単語難易度の辞書を自動で構築した.

日本語の難易度推定の既存研究では, 単語レベルの難易度推定や, 文章単位のリーダビリティとしての難易度推定が行われているが, 非母語話者の記述練習などの目的では文単位の評価が望ましいと考えた. 特に, 文章単位の難易度推定では, 使用している特徴量として, 主に文章の単純な統計量のみを考慮しており, 文法や文脈などの情報を十分に加味したものは存在していない. また, 非母語話者向けの文章の難易度判定の既存研究では, 独自の尺度で難易度を推定している. これらは, 学習者が自身の能力評価に用いる日本語能力試験（JLPT）の習熟度基準とは直接的な対応関係がなく, 難易度の評価結果が直感的ではない. さらに, 日本語文の難易度評価や教材自動作成のための言語資源で, 最新の JLPT の習熟度基準に準拠しているものはそもそも存在しておらず, 旧習熟度基準のラベル付きコーパスについても公開されているものは存在しない. そこで, 本研究では, 単文の難易度を JLPT の習熟度基準に準拠した難易度で推定するため, JLPT の過去の試験問題や学習参

表 1　JLPT の習熟度基準

習熟度	基準
N1	様々な話題の, 内容に深みのある読み物の内容や詳細な表現意図を理解できる. 自分の関心のある分野のレポートを記述できる.
N2	新聞や雑誌の記事、解説, 平易な評論など, 論旨が明解な文章が理解できる. 一般的な話題に関する読み物の話の流れや表現意図を理解できる. 理由を述べながら意見を述べたり, 学校, ホテルに問い合わせなどの連絡を記述できる.
N3	日常的な話題についての具体的な内容を表す文を理解できる. 新聞の見出しなどから情報の概要を理解できる. 難易度が高い文章でも, 言い換えられると要旨を理解できる. 知人に感謝・謝罪を伝える手紙を記述できる.
N4	基本的な語彙や漢字で書かれた, 日常生活で身近な話題の文章を理解できる. 日常の要件を伝える簡単なメモを記述できる. 依頼や誘いなどの簡単な文章が記述できる.
N5	平仮名, カタカナ, 基本的な漢字を理解できる. 定型的な語句, 文, 文章を理解できる. 書類に名前や国名などを記述できる. 簡単な自己紹介や短いお礼を記述できる.

考書からラベル付きコーパスを作成する. 作成したコーパスをもとに, 文の難易度推定器の作成する.

2　JLPT の習熟度基準

本研究では, コーパスから日本語文の難易度を推定する. 難易度推定結果を直感的に理解しやすくするため, 非母語話者の日本語運用能力を図る試験として広く用いられている JLPT の習熟度基準に準拠するよう, JLPT の過去試験や学習参考書から文の難易度ラベル付きコーパスを作成する. そこで, まず JLPT の習熟度基準について説明する.

JLPT では, 日本語の運用能力を上級, 中級, 初級の 3 段階で定義している. 上級の能力は, 現実の生活の幅広い場面で使われる日本語の理解ができることを示す. 初級の能力は, 学習者が日本語教室内で学ぶレベルの日本語の理解ができることを示す. 中級は, 初級から上級へ過渡的な級である. また, 3 つ級を分割し, N1 から N5 の 5 段階で能力を評価している.

表 1 に, JLPT の定義する習熟度基準を示した. 上級の N1, N2 では, 新聞や論説をはじめとする高度な文章の読解力と, 自分の意見の記述力など, 高度な日本語運用能力を評価する. また, 初級の N4, N5 では, 日本語の初学者の学習状況を評価する. N1 または超級と呼ばれる段階では, 全ての難解なテキストが N1 と表現されるため, N1 の内部でも難易度に大きな差があるという課題がある.

表 2　2012 年過去問の統計情報

難易度	サンプル数	文の平均文字数	文の漢字比率
N1	61	27.9	0.317
N2	169	27.3	0.296
N3	203	25.5	0.275
N4	177	19.6	0.136
N5	118	15.3	0.061

3　JLPT の習熟度基準に基いたデータ作成

JLPT の習熟度基準に基づいた文の難易度を予測するため, JLPT の過去の公式試験問題と, 学習参考書を購入し, スキャンを行った. スキャンした画像データから, OCR[1)]で文字を抽出し, 抽出した文字から文を構成した. 主に, 語彙・文法の問題や解説文に単文が多く含まれていたため, 本研究ではそれらの文を抽出した.

OCR による抽出後の後処理として, 文のルビを削除し, 文分割を行った. 文分割において, 括弧"「", "」"で囲まれた文章は, 単文として扱った. また, 試験問題の冒頭では, 会話者の名前や問題番号があるが, それらは取り除いた. 参考書について, 外国人向けの参考書であったので, 外国語を取り除く後処理も行なっている.

3.1　JLPT の試験問題を用いた評価データ

JLPT の習熟度基準に準拠した難易度を評価するため, 販売されている 2012 年の試験問題集 [15] から問題文の抽出を行った. 試験問題のうち, 語彙, 文法, 読解問題の問題文や解答選択肢の文や, 読解問題の文章から文を抽出した. 初級 (N4, N5) はひらがなや簡単な漢字のみの短い文が多くあり, 上級 (N1, N2) では漢字などが自然に用いられた比較的長い文が多く存在している. 表 2 より, 文内の漢字比率について, 特に N5 には, ほとんど漢字が含まれていない傾向が確認できる. また, 初級の問題は語彙, 文法などの学習中の知識を問う問題が多いため, 表 2 のように上級と比べて多くのサンプルを取得できた. 一方で, 上級は語彙・文法ではなく読解力を問う問題が多いため, 抽出できたサンプルが少ない.

3.2　試験参考書を用いた訓練データ

データを拡張するため, 試験問題に比べて豊富に存在する JLPT の学習参考書から 3.1 節と同様に文の抽出を行った. 学習参考書は, 上級では文法項目を

1)　https://clova.line.me/clova-ocr/

表 3　訓練データの統計情報

難易度	サンプル数	文の平均文字数	文の漢字比率
N1	348	22.9	0.263
N2	380	18.0	0.244
N3	437	18.1	0.203
N4	413	17.0	0.193
N5	730	23.6	0.087

簡潔に説明するための短めの例文が多く含まれていた. また初級では, 学習者の漢字の学習促進のため, 簡単な漢字は文内でルビつきの状態で用いられていた. 表 3 でも, 表 2 と比較して, 全体的に文長が短く, 中〜初級の漢字の使用比率が高くなっている.

4　機械学習による難易度推定

4.1　実験設定

学習参考書より作成したデータセットを用いて, 文の難易度の推定器を作成する. 推定器を JLPT の過去の試験問題より作成した評価データで評価する.

モデルへの入力として, 文から作成したベクトルと, 文章の難易度推定で用いられている, 文長と漢字の比率を与える. まず, 文からベクトルを作成する手法を次に示す.

1. BERT Emb: 日本語の BERT[2] の max-pooling
2. TF-IDF: 訓練データで作成した TF-IDF
3. BM25: 訓練データで作成した Okapi BM25[16]
4. jRead: jReadability [9] で用いられた特徴量

ある単語の出現頻度は, そのまま使うと文書の長さに従った値になる. この問題を緩和するため, 重要度の高い単語に重み付けを行える TF-IDF や BM25 による文のベクトル化を行う. また, 文脈を考慮するため BERT による max-pooling によるによる文のベクトル化を行う. また, 既存研究との比較として, jReadability.net で用いられている, 文長, 漢語の比率, 和語の比率, 動詞の比率, 助詞の比率からなる, 5 つの文の統計量を文のベクトルとする jRead を用いた比較実験も行う. jRead の比較実験では, 文長, 漢字比率は jRead に包含されているため結合しない. また, それぞれの入力について, 単語難易度情報の有無による性能差を検証するため, 単語難易度情報を入力データに結合した実験 (+ JEV) も行う. 日本語教育語彙表[3] を用いて, 入力文内の単語について, 各単語難易度の出現頻度を計算した 6 次元の頻度ベクトルを単語難易度情報として用いる.

それぞれ, 作成した問題文の特徴量から JLPT の N1 から N5 の難易度を推定するように, LightGBM[4] の推定モデルの訓練を行う. 難易度は, 離散値なので判別 (clf) も行うが, 同様に大小関係があるため, 平均二乗誤差を最小化する回帰 (reg) としても訓練を行う. LightGBM を用いたモデルは, Optuna[5] を用いて訓練データの交差検証によるチューニングを行う. また, 既存研究との比較のため, jRead については, 線形回帰モデル (lin reg) も実装する.

表 4　2012 年度過去試験のテキストによるモデルの評価

	MAE	Pearson	Spearman	Acc	F1
BERT Emb + JEV (reg)	**0.548**	0.729	0.749	0.511	0.477
BERT Emb + JEV (clf)	0.738	0.724	0.738	0.448	0.446
BERT Emb (reg)	0.549	0.729	0.743	**0.515**	**0.496**
BERT Emb (clf)	0.734	0.713	0.731	0.456	0.456
BM25 + JEV (reg)	0.588	0.723	0.741	0.464	0.422
BM25 + JEV (clf)	0.705	**0.747**	**0.755**	0.455	0.455
BM25 (reg)	0.604	0.720	0.737	0.446	0.397
BM25 (clf)	0.809	0.687	0.696	0.394	0.382
TF-IDF + JEV (reg)	0.584	0.729	0.746	0.466	0.433
TF-IDF + JEV (clf)	0.706	0.741	0.751	0.448	0.449
TF-IDF (reg)	0.628	0.692	0.705	0.435	0.409
TF-IDF (clf)	0.859	0.670	0.682	0.374	0.359
jRead + JEV (reg)	0.661	0.671	0.680	0.413	0.353
jRead + JEV (clf)	0.868	0.650	0.659	0.387	0.369
jRead (reg)	0.827	0.457	0.458	0.331	0.283
jRead (clf)	1.212	0.424	0.425	0.290	0.268
jRead (lin reg)	0.916	0.440	0.463	0.291	0.273
jReadability.net[9]	-	0.440	0.455	-	-

4.2　実験結果

作成したモデルの評価指標として, 平均絶対誤差 (MAE), ピアソンの相関係数 (Pearson) , スピアマンの順位相関係数 (Spearman), 難易度予測の正解率 (Acc), F1 スコアを表 4 に示す. 回帰モデル (reg) の実験結果については, 回帰結果を四捨五入した値を擬似的なラベルとみなして, 擬似ラベルとの評価指標を計算した. また, 比較のため, 公開されている文章の難易度評価システムである jReadability.net[6] [9] で, リーダビリティスコアを測定し, リーダビリティスコアと難易度との相関係数も計算した. jReadability.net について, 相関係数以外の指標は尺度が異なるため表記していない.

文章ではなく文単位の予測にとって, jRead によるモデルや jReadability.net の回帰式では十分ではない

2) https://huggingface.co/cl-tohoku/bert-base-japanese-v2
3) http://jhlee.sakura.ne.jp/JEV/
4) https://github.com/microsoft/LightGBM
5) https://github.com/optuna/optuna
6) https://jreadability.net/sys/

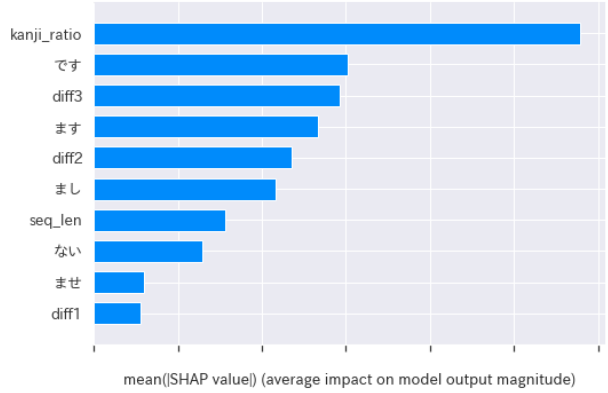

図 1　TF-IDF + JEV (reg) による shap 値

表 5　BERT Emb + JEV (reg) の混同行列

	N1	N2	N3	N4	N5
N1	0	38	16	7	0
N2	0	129	37	3	0
N3	0	87	98	15	3
N4	0	3	48	111	15
N5	0	0	4	80	34

表 6　BERT Emb + JEV (clf) の混同行列

	N1	N2	N3	N4	N5
N1	30	15	8	4	4
N2	78	58	26	6	1
N3	66	34	63	31	9
N4	4	11	12	66	84
N5	0	1	3	5	109

ことが示された. また, 同一の特徴量での reg と clf の比較として, 自動評価では reg が clf より精度が高い傾向が見られた. また, 単語難易度の情報は文難易度の予測においても重要であることが確認された.

4.3　難易度推定モデルの分析

入力文において, どの単語や統計情報が予測にとって重要かを分析するため, TF-IDF + JEV で作成した回帰モデルに対して, SHAP[17] による特徴量の寄与率を分析した. 図 1 より, 漢字や文字数などの単純な統計量は文章の難易度推定と同様に, 文の難易度推定においてもある程度重要であることを確認した. また, 文の難易度推定にとって, 単語の難易度情報も重要であることも確認した. また, "です", "ます" といった丁寧語の文末表現は, 初級のような教科書上のテキストに出やすい傾向があった. 丁寧語の文末表現は, 初級の推定において活用される傾向が高かったため, 敬語の文末表現の寄与率が高まる傾向が現れたと考えられる.

表 5 では, MAE において最も精度の高かった, BERT Emb + JEV を用いた回帰モデルの出力結果を四捨五入した擬似的なラベルの混同行列を示している. 表 5 のように, 回帰タスクとして訓練を行った結果, N1 の推定がほぼ適切には行えていないことが確認された. 一方で, 表 6 で示した, BERT Emb + JEV を用いた判別モデルの混同行列では, N1 や N5 など難易度の両端の予測については回帰に比べて高精度だが, 全体的には誤判別が増加している.

表 7　BERT Emb + JEV を用いた回帰モデルの誤判別例

	正解	予測	入力文
1	N1	N4	待っておいでになります
2	N1	N2	そして、日本人のやきものに対する思いとか愛着は、食器のみならず、種類の豊富さにあらわれているといってもいいでしょう。
3	N2	N4	いつも「ジミック」のプリンターをご愛用いただき、ありがとうございます。
4	N3	N4	明日 (30 日) の約束ですが、会議に出なければならなくなりました。
5	N4	N2	約束を明日に変えられるかどうか
6	N4	N3	自転車やオートバイは、公園の入り口にとめてください。
7	N4	N5	ふねでにもつをおくります。
8	N5	N4	このまちにはゆうめいなビルがあります。

表 7 に, BERT Emb + JEV を用いた回帰モデルの誤判別の事例を示している. 上級や中級を初級と誤判別していた事例の多くは, 表 7 の例 1, 3, 4 のように, 手紙やパンフレットなどの, 挨拶や簡単な伝達事項の連絡文など, 定型的な表現の文であった. 定型表現は, 2 章で示したように, JLPT の初級の表現であるが, 前後の文脈の兼ね合いで上級の問題にも一定程度含まれる. 定型表現が上級にも含まれる影響で, 例 5 のように初級のテキストを上級に誤判別している可能性がある. また, 例 2, 6, 7, 8 のように, 隣接する難易度での推定ミスが多く存在している.

5　おわりに

本研究では, 非母語話者の日本語の学習支援に向けて, 文の難易度を JLPT の習熟度基準に従って推定するため, JLPT の試験問題や学習参考書から, 難易度ラベル付きコーパスを作成した. さらに, 作成したコーパスから, 文脈情報を加味できる特徴量を用いた文難易度の推定器を作成した. 推定器の評価を行った結果, 難易度ラベルと高い相関を持つ推定が行えることを確認した. さらに, 誤判別例や混同行列を用いたエラー分析により, 文難易度の推定に関して, 本質的な課題として基礎的な文はある程度どの難易度にも出現しうることや, 文難易度を予測するためのコーパス設計上の課題を示した. 今後の課題として, さらなるデータの拡張を行うことで, 基礎的な文の影響を低減させることや, 順序を考慮したモデルの活用などによる推定モデルの改善や, N1 以上の習熟度基準の定義やその推定に取り組む.

参考文献

[1] 出入国在留管理庁. 在留支援のためのやさしい日本語ガイドライン, 2020. https://www.moj.go.jp/isa/support/portal/plainjapanese_guideline.html.

[2] 文化庁. 令和2年度国内の日本語教育の概要, 2021. https://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/nihongokyoiku_jittai/r02/.

[3] J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. **Technical report, Defence Technical Information Center Document**, pp. 8–75, 1975.

[4] 林正頼, 石井康毅, 高村大也, 奥村学, 投野由紀夫. 英語学習者の英作文から CEFR レベル別基準特性の特定. 言語処理学会 第22回年次大会, pp. 781–784, 2016.

[5] 林正頼, 笹野遼平, 高村大也, 奥村学. 誤り傾向と文の容認性に着目した英作文のレベル判定. 情報処理学会研究報告, No. 7, pp. 1–7, 2016.

[6] 川村よし子, 北村達也. 日本語学習者のための文章と難易度判定システムの構築と運用実験. **Journal CAJLE**, No. 14, pp. 18–30, 2013.

[7] Yuka Tateisi, Yoshihiko Ono, and Hisao Yamada. A Computer Readability Formula of Japanese Texts for Machine Scoring. In **Proceedings of the 12th Conference on Computational Linguistics**, pp. 649–654, 1988.

[8] 柴崎秀子. リーダビリティー研究と「やさしい日本語」. 日本語教育, Vol. 158, pp. 49–65, 2014.

[9] 李在鎬. 日本語教育のための文章難易度に関する研究. 早稲田日本語教育学, No. 21, pp. 1–16, 2016.

[10] Jun Liu and Yuji Matsumoto. Sentence Complexity Estimation for Chinese-speaking Learners of Japanese. In **Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation**, pp. 296–302, 2017.

[11] Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In **Proceedings of the Sixth International Conference on Language Resources and Evaluation**, pp. 654–660, 2008.

[12] 佐藤理史. 均衡コーパスを規範とするテキスト難易度測定. 情報処理学会論文誌, Vol. 52, No. 4, pp. 1777–1789, 2011.

[13] Yuriko Sunakawa, Jae-Ho Lee, and Mari Takahara. The Construction of a Database to Support the Compilation of Japanese Learners' Dictionaries. **Acta Linguistica Asiatica**, Vol. 2, No. 2, pp. 97–115, 2012.

[14] Daiki Nishihara and Tomoyuki Kajiwara. Word Complexity Estimation for Japanese Lexical Simplification. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 3114–3120, 2020.

[15] 日本語能力試験公式ウェブサイト. 『日本語能力試験公式問題集』 | 日本語能力試験 JLPT, 2012. https://www.jlpt.jp/samples/sampleindex.html.

[16] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In **Proceedings of The Third Text REtrieval Conference**, pp. 109–126, 1994.

[17] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In **Advances in Neural Information Processing Systems**, pp. 4768–4777, 2017.

A　データ作成に使用した書籍情報

A.1　評価データ

表 8　評価データの例

難易度	例文
N1	そして、日本人のやきものに対する思いとか愛着は、食器のみならず、種類の豊富さにあらわれているといってもいいでしょう。
N2	マスコミで毎日のように環境問題が取り上げられているが、本当に「環境問題」と言っていいのだろうか。
N3	実際に印刷した紙がプリンターのところに置いてありますからそれを修理の人に見せて説明してください
N4	でんしゃの中でさわがないでください。
N5	いりぐちはあちらです。

表 9　評価データ作成箇所

試験問題	サンプル取得場所
N1	言語知識: 問 1-3, 問 5 の問題文 言語知識: 問 4 の解答選択肢の文 言語知識: 問 8 (1), (2), 問 9-12 の問題文章と解答選択肢の文 言語知識: 問 8 (2) の解答選択肢の文
N2	言語知識: 問 1-5, 7, の問題文 言語知識: 問 6 の解答選択肢の文 言語知識: 問 10-13 の問題文章と解答選択肢の文
N3	言語知識 (文字・語彙): 問 1-4 の問題文 言語知識 (文字・語彙): 問 5 の解答選択肢の文 言語知識 (文法・読解): 問 1 の問題文 言語知識 (文法・読解): 問 4 (1) (3) (4), 問題 5, 6 の問題文章と解答選択肢の文
N4	言語知識 (文字・語彙): もんだい 1,2,3 の問題文 言語知識 (文字・語彙): もんだい 4 の問題文と解答選択肢の文 言語知識 (文字・語彙): もんだい 5 の解答選択肢の文 言語知識 (文法・読解): もんだい 1 の問題文 言語知識 (文法・読解): もんだい 4 (1), (3), (4), もんだい 5 の問題文と解答選択肢の文 言語知識 (文法・読解): もんだい 4 (2) の解答選択肢の文
N5	言語知識 (文字・語彙): もんだい 1-3 の問題文 言語知識 (文字・語彙): もんだい 4 の解答選択肢の文 言語知識 (文法・読解): もんだい 1 の問題文 言語知識 (文法・読解): もんだい 4 (1), (3), もんだい 5 の問題文章と解答選択肢の文

評価データの作成のため, JLPT の 2012 年の公式過去試験問題集を購入した.

- 国際交流基金, 日本国際教育支援協会. 日本語能力試験 公式問題集 N1. 凡人社. 2012.
- 国際交流基金, 日本国際教育支援協会. 日本語能力試験 公式問題集 N2. 凡人社. 2012.
- 国際交流基金, 日本国際教育支援協会. 日本語能力試験 公式問題集 N3. 凡人社. 2012.
- 国際交流基金, 日本国際教育支援協会. 日本語能力試験 公式問題集 N4. 凡人社. 2012.
- 国際交流基金, 日本国際教育支援協会. 日本語能力試験 公式問題集 N5. 凡人社. 2012.

また, これらの書籍に掲載されている実際の試験問題は, 日本語能力試験公式 ウェブサイト[7] より取得できる. 実際に評価データの作成に使用した箇所を表 9 に示す. 抽出したデータのうち, 文内に空欄が存在したが, 空欄は空欄を埋めずに使用した. 作成した訓練データのサンプルを表 8 に示す.

A.2　訓練データ

表 10　訓練データの例

難易度	例文
N1	彼の作文は小さい間違いこそあれど、よく書けている。
N2	風邪気味でこの季節は熱っぽいんだ。
N3	けがをされると困るからダメよ。
N4	兄は日本で働いています。
N5	お国はどちらですか。

JLPT の習熟度基準に準拠した学習参考書から訓練データを作成した.

- 佐々木 仁子, 松本 紀子. 日本語総まとめ N1 語彙 英語・ベトナム語版. アスク出版, 2019.
- 佐々木 仁子, 松本 紀子. 日本語総まとめ N1 文法 英語・ベトナム語版. アスク出版, 2019.
- 佐々木 仁子, 松本 紀子. 日本語総まとめ N1 読解 英語・ベトナム語版. アスク出版, 2019.
- 佐々木 仁子, 松本 紀子. 日本語総まとめ N2 語彙 英語・ベトナム語版. アスク出版, 2018.
- 佐々木 仁子, 松本 紀子. 日本語総まとめ N2 文法 英語・ベトナム語版. アスク出版, 2018.
- 佐々木 仁子, 松本 紀子. 日本語総まとめ N2 読解 英語・ベトナム語版. アスク出版, 2018.
- 佐々木 仁子, 松本 紀子. 日本語総まとめ N3 語彙 英語・ベトナム語版. アスク出版, 2018.
- 佐々木 仁子, 松本 紀子. 日本語総まとめ N3 文法 英語・ベトナム語版. アスク出版, 2018.
- 佐々木 仁子, 松本 紀子. 日本語総まとめ N3 読解 英語・ベトナム語版. アスク出版, 2018.
- 佐々木 仁子, 松本 紀子. 日本語総まとめ N4 漢字・ことば 英語・ベトナム語版. アスク出版, 2018.
- 佐々木 仁子, 松本 紀子. 日本語総まとめ N4 文法・読解・聴解 英語・ベトナム語版. アスク出版, 2018.
- 佐々木 仁子, 松本 紀子. 日本語総まとめ N5 かんじ・ことば・ぶんぽう・読む・聞く 英語・ベトナム語版. アスク出版, 2018.

後処理の都合上, 主に文法の解説のための例文を抽出して使用した. 具体的には, N1,2,3 については, 文法の教科書の全ての解説例文を抽出している. また, N4 は, 読解や文法の使い方の説明文と, 読解問題の文章を抽出した. N5 については, 書籍前半の語彙と, 「聞く」パート以外の, 全ての語彙・文法などの使用例と, 読解問題の文章を抽出している. 作成した訓練データのサンプルを表 10 に示す.

7)　https://www.jlpt.jp/samples/sampleindex.html