修士論文

中国語母語話者を対象とした日本語単語の難易度推定

LIN Miaoyu

主指導教員 白井 清昭

北陸先端科学技術大学院大学
先端科学技術研究科
(情報科学)

令和 4 年 3 月

Abstract

Identification of difficulty of words plays an essential role in teaching Japanese. There are many situations where the difficulty level of a word is considered, such as prioritizing the words to be taught to learners or not using difficult words for beginners. In general, the difficulty of words depends on the learner's background. For example, since learners whose mother tongue is Chinese know Chinese characters, they can learn Japanese words written in Chinese characters more easily than learners whose mother tongue is a language that does not use Chinese characters. However, previous studies on the difficulty of Japanese words did not consider the difference in learners' native languages.

This study aims at estimating the difficulty of Japanese words for native Chinese speakers. We focus on estimation of the difficulty level when learners study a word for the first time, and also focus on only Japanese words written in Chinese characters (Kanji). In order to measure the difficulty level of Japanese words for native Chinese speakers, for a given Japanese word, a Chinese word written in the same Chinese characters is searched first. When it is found, we measure how similar the senses of Japanese and Chinese words are. The more similar the senses are, the lower the difficulty level is regarded to be. The similarity of word senses is measured by the alignment of the senses in a Japanese dictionary and a Chinese dictionary. In addition, more fine-grained difficulty levels are identified by checking whether the glyphs of the Japanese and Chinese words are completely identical or not.

In this study, we define the difficulty level of Japanese words as follows. Based on the classification of Sino-Japanese words in the past document, four difficulty classes are defined in ascending order of difficulty: S (all senses in Chinese and Japanese words are the same), O (some senses in Chinese and Japanese words are overlapped), D (senses in Chinese and Japanese are totally different), and N (the same word does not exist in Chinese). In addition, even when a Japanese Kanji character and Chinese one are the same, they are sometimes represented by different glyphs. Therefore, the classes S, O, and D are subdivided into X-1 when the glyphs of Japanese and Chinese words are the same, and X-2 when they are different, where X stands for S, O or D.

The classification of S, O, D, and N for Japanese Kanji words is carried out as follows. First of all, in a Chinese dictionary, we search the same Chinese word as a given Japanese Kanji word. If the word is not found, we convert the Japanese Kanji character to Simplified Chinese using the Chinese-Japanese Kanji mapping table, then search again. When the word is not found even after Kanji conversion, it is classified as N. Otherwise, we extract definition sentences of the senses of the words from the Japanese dictionary and two Chinese dictionaries. The Iwanami

Japanese Dictionary is used as the Japanese dictionary, while the Hakusuisha Chinese Dictionary and the Contemporary Chinese Dictionary are used as Chinese dictionaries. The senses are written in Chinese in the Contemporary Chinese Dictionary, so they are translated into Japanese using the Baidu Translation API. Next, we calculate the similarity between the word senses in the Japanese and the Chinese dictionary, and align similar word senses. A word sense is represented as a vector, which is the average vector of the distributional representations of words in a definition sentence. The similarity between the word senses in the two dictionaries is defined as the cosine similarity of the two word sense vectors. Among all the combinations of Japanese and Chinese senses, the pair of senses with the highest similarity is aligned first. After removing the aligned word senses from the two dictionaries, the same process is repeated for the rest of the word senses. However, when the similarity between the senses is less than a threshold T_m, that is, when the sense similarity is not sufficiently large, we do not consider that the two senses have the same meaning and terminate the alignment of the senses. After the alignment of word senses is completed, when none of the pairs of word senses can be aligned, it is judged as D. When not all but some of word senses are aligned, it is judged as O. When all of the word senses are aligned, it is judged as S. We perform the sense alignment for each of Chinese dictionaries and Japanese dictionary. If the results of the two dictionaries are different, we calculate the score of the word sense alignment, which is the average of the similarity between the aligned word senses, and choose the result with the highest score. When the Japanese word has two or more parts-of-speech (POSs), the sense sets are subdivided according to its POS, then the sense alignment is carried out for each subset of the senses. Finally, the difficulty class is determined as S-1, O-1 or D-1 when the glyphs of the Japanese and Chinese word are the same, while S-2, O-2, or D-2 when they are not. In addition, we propose another method to use Word Mover's Distance to measure the similarity between the senses. Also, we propose another algorithm that considers many-to-many alignment between the senses.

The proposed method was evaluated from two points of view. First, we evaluated the performance of the proposed method in identifying the difficulty level. In the experiment, only S, O and D were considered as the difficulty level, since the discrimination between N and others, and between X-1 and X-2 was obvious. A test data consisting of 279 Japanese words with its gold difficulty level was manually constructed. Then the accuracy of the classification of the difficulty level was measured on this test data. As a result, the accuracy was 0.763 when the threshold T_m was set to 0.196 by heuristics. Although there was still room for improvement, it was confirmed that the proposed method could identify the difficulty level of Japanese words with the reasonably high accuracy.

Second, the validity of the proposed word difficulty classes was verified by a questionnaire survey of native speakers of Chinese. We extracted 5 words for each class S, O, D and N from the test data and asked 22 native Chinese speakers to rate the difficulty of these 20 Japanese words on a 5-point scale by a questionnaire. The rank correlation coefficient between the difficulty of the proposed method and the average of the rating evaluated by the Chinese native speakers in the questionnaire survey were calculated. The results showed that for beginners, the correlation coefficient was 0.788 with a p-value of 0.00004. In other words, a strong correlation between them was found, which was statistically significant. Therefore, it is found that the proposed framework is appropriate to measure the difficulty of Japanese words for beginners.

The main contribution of this thesis was to establish the method for automatically estimating the difficulty of Japanese words with the high accuracy, taking into account the characteristics of native Chinese speakers who know Kanji characters.

概要

日本語教育において，単語の難易度は重要な役割を果たす．学習者に教える単語の優先順位を決めたり，初学者に対して難しい単語の使用を避けたりするなど，単語の難易度を必要とする場面は多い．一般に，単語の難易度は日本語学習者の背景知識に依存すると考えられる．例えば，中国語を母語とする学習者は漢字を知っているので，漢字を使用しない言語を母語とする学習者よりも漢字表記の単語を学習しやすいと考えられる．ところが，日本語単語の難易度に関する先行研究では，学習者の母語の違いは考慮されていなかった．

本研究は，中国語母語話者を対象とした日本語単語の難易度を推定することを目的とする．ただし，単語を初めて勉強したときの難易度を研究の対象とし，難易度を推定する単語は漢字で表記された単語に限る．漢字を知っている中国語母語話者にとっての日本語の単語の難易度を推定するために，漢字で表記された日本語単語に対して，それと同じ中国語の単語があるかをチェックする．同じ単語があるとき，日本語の意味と中国語の意味がどれだけ似ているかを測り，似ている単語ほど難易度が低いと推定する．単語の意味の類似度は日中の単語辞書における語義の対応付けによって測る．また，日本語単語と中国語単語の漢字表記が完全に一致しているかどうかによって難易度をさらに細分化する．

まず，単語の難易度を以下のように定義する．文化庁による資料「中国語と対応する漢語」における単語の分類を基に，難易度が低い順に，大きく S(日中における意味が全て同じ)，O(日中における意味が一部重なっている)，D(日中における意味が著しく異なる)，N(日本語と同じ単語が中国語に存在しない) の 4 つのクラスに分ける．また，日本語と中国語では同じ漢字が異なる字体で表されることがある．そこで，クラス S, O, D については，日本語単語と中国語単語の漢字表記が同じときは X-1，異なるときは X-2 のように細分化する．

漢字表記語の S, O, D, N の分類は以下のように行う．まず，日本語の漢字表記語と同じ単語が中国語の辞書にあるかを検索する．もし見つからない場合，日中漢字マッピングテーブルを参照し，中国語の漢字に変換してから検索する．漢字を変換しても見つからない場合，N と分類する．それ以外は，日本語辞書から日本語単語の語義を，中国語辞書から中国語単語の語義を抽出する．また，語義とともにその語釈文も抽出する．日本語の辞書として岩波国語辞典を，中国語の辞書として白水社中国語辞典と現代漢語詞典を用いる．現代漢語詞典では語釈文は中国語で書かれているが，Baidu 通用翻訳 API を用いて日本語に翻訳する．次に，日中の辞書の語義の類似度を計算し，似ている語義の対応付けを行う．語義のベクトルは，その語義の語釈文に含まれる単語の分散表現の平均ベクトルとする．日中の語義の類似度は，2 つの語義のベクトルのコサイン類似度で算出する．日本語の語義と中国語の語義の全ての組み合わせのうち，類似度が最も高くなる語義の組をまず対応付ける．対応付けられた語義の組を除き，残りの語義の組について同じ処理を繰り返す．ただし，語義の類似度が閾値 T_m より小さいとき，つまり語義間の類似度が十分に大きくないとき，二つの語義は同じ意味を持つとはみなさ

ず，語義の対応付けを終了する．語義の対応付けの終了後，対応付けできる語義の組が 1 つも見つからないときは D と判定し，対応付けできる語義の組はあるが全ての語義について対応付けできないときは O と判定し，全ての語義について対応付けができたときは S と判定する．上記の判定を中国語の辞書として白水社中国語辞典と現代漢語詞典を用いたときのそれぞれについて行い，2 つの辞書による判定結果が異なる場合は，語義の対応付けのスコア (対応付けられた語義間の類似度の平均値) を算出し，大きい方の判定結果を採用する．また，日本語単語が多品詞語のとき，品詞毎に語義を分けた上で難易度の判定を行う．最後に，日本語単語と中国語単語の漢字表記が一致しているときは S-1, O-1, D-1 のいずれかを，一致していないときは S-2, O-2, D-2 のいずれかを難易度クラスとする．この他に，語義間の類似度として Word Mover's Distance を用いる手法や，語義の対応付けを行う際に多対多の対応付けを考慮する手法も提案する．

提案手法を 2 つの観点で評価した．一つ目は，提案手法による難易度判定の性能を評価した．N とそれ以外の区別，X-1 と X-2 の区別は自明なので，ここでは S, O, D の判定を行った．人手で正解の難易度を付与した 279 語の漢字表記語をテストデータとし，提案手法による難易度判定の正解率を調べた．その結果，ヒューリスティクスによって閾値 T_m を 0.196 と設定したとき，難易度判定の正解率は 0.763 となった．改善の余地はあるものの，提案手法による難易度判定はある程度高い水準にあると言える．二つ目は，本研究で提案する日本語単語難易度の尺度が，実際の中国語母語話者から見て妥当であるかを検証した．テストデータから S, O, D, N クラスの単語をそれぞれ 5 語抽出した．22 人の中国語母語話者に対してアンケート調査を行い，これらの 20 語の日本語単語の難易度を 5 段階で評価してもらった．提案手法による単語難易度とアンケート調査の回答結果による単語難易度の順位相関係数を求めた．その結果，初学者に対しては 1% の有意水準で両者に相関があり，提案手法の単語難易度の尺度の妥当性が確認された．

本論文の主たる貢献は漢字を知っているという中国語母語話者の特性を考慮した日本語単語の難易度を提案し，それを高精度で自動的に推定する手法を確立した点にある．

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

日本語教育において，日本語単語の難易度は重要な役割を果たす．例えば，日本語教師が講義する際には，初学者でも講義の内容を理解できるように，教える単語の優先順位を決めたり，難しい単語の使用を避けている．また，初学者には理解が難しい日本語のテキストがあったとき，これを学習者の習熟度に応じて理解しやすいテキストに書き換えるには，日本語の単語の難易度をあらかじめ策定する必要がある．

一般に，単語の難易度は日本語学習者の背景知識に依存すると考えられる．例えば，中国語を母語とする学習者は漢字を知っているので，漢字で単語の意味が推測できることがあり，そのような単語の難易度は低くなると考えられる．したがって，中国語母語話者は漢字を使用しない言語を母語とする学習者よりも漢字表記の単語を学習しやすいと考えられる．外国人を対象とした日本語単語の難易度判定基準として，日本語能力試験の受験級が広く知られている．しかし，この難易度の基準では学習者の母語の違いは考慮されていなかった．日本語能力試験で難易度が高いと思われる単語であっても，漢字で表記する語については，中国語を母語とする学習者は読み方が分からなくても漢字で意味が推測できる場合は多々ある．沖森らが『新選国語辞典 (第九版)』の見出し語 76,536 語を調べた結果，漢語が占める割合は 49.4%，和語は 32.2%であったという [11]．すなわち，漢字で表記された漢語は半数近くを占めている．中国語母語話者は，正しく推測できるかどうかはともかく，半数近くの日本語単語は漢字である程度意味を推測できると言える．

国際交流基金の「海外の日本語教育の現状　2018 年度日本語教育機関調査」によれば，中国の日本語学習者数は 100 万人を超え，日本語学習者の最も多い国である．台湾やシンガポールなどほかの国における中国語を母語とする日本語学習者も加えると，中国語母語話者の日本語学習者の数は同調査で報告されている数よりもさらに多くなる．日本語教育の現場において学習者の背景知識の違いを考慮するときには，中国語を母語とする学習者は無視できない存在である．

以上から，日本語教育の場面で用いる単語の難易度は学習者の母語を考慮して定義するべきであり，特に学習者の多さから中国語を母語とする学習者から見た日本語単語の難易度を推定することが重要である，しかしながら，日本語単語の難易度に関する先行研究の多くは，日本語母語話者，またはあらゆる外国人学習

者を対象とし，学習者の母語の違いは考慮されていなかった．特に，漢字を知っている中国語母語話者にとっての日本語単語の難易度は，汎用的な単語の難易度を流用するだけでは不十分である．

1.2 目的

本研究は，日本語の辞書と中国語の辞書における語義の違いや漢字表記の違いに基づき，中国語母語話者を対象とした日本語単語の難易度を推定することを目的とする．ここでの単語の難易度は，単語を習得するときの難しさ，すなわち単語を最初に学ぶ際に意味を推測する難しさを指すものとする．また，難易度を推定する日本語単語は漢字で表記された単語に限定する．漢字を知っている中国語母語話者にとっての日本語の単語の難易度を推定するために，漢字表記語に対して，これと同じ中国語の単語があるかをチェックする．同じ単語があるとき，日本語の意味と中国語の意味がどれだけ似ているかを測り，似ている単語ほど難易度が低いと推定する．単語の意味の類似度は日中の単語辞書における語義の対応付けによって測る．また，日本語単語と中国語単語の漢字表記が完全に一致しているかどうかによって難易度をさらに細分化する．

本論文では，単語の難易度を推定するタスクの正解率を測る実験を行い，提案手法の有効性を評価する．さらに，日本語学習者に単語の難易度に関するアンケート調査を行い，被験者が感じる単語の難易度が本研究で提案する単語の難易度とどれだけ一致しているかを検証する，

1.3 本論文の構成

本論文の構成は以下の通りである．第 2 章では，本論文の関連研究を紹介する．第 3 章では，提案手法について説明する．第 4 章では，提案手法の評価実験とその結果について述べる．最後に，第 5 章では，本論文のまとめと今後の課題について述べる．

第2章　関連研究

本章では，本論文に関連する研究について述べる．2.1 節では，日本語の難易度判定に関する研究を紹介する．2.2 節では，日本語単語の漢字処理おける母語の影響に関する研究を紹介する．2.3 節では，テキストの類似度に関する研究について述べる．本論文の提案手法では，単語の難易度を判定するとき，テキストの類似度を測る必要があるため，それに関連する研究を紹介する．最後に，2.4 節では本研究と関連研究の違いについて論じる．

2.1　日本語単語の難易度判定に関する研究

ここでは日本語単語の難易度を判定する先行研究を 2 つに分けて紹介する．2.1.1 項では日本語母語話者を対象とした日本語単語の難易度判定に関する研究を，2.1.2 項では外国人を対象とした日本語単語の難易度判定に関する研究を紹介する．

2.1.1　日本語母語話者を対象とした日本語単語の難易度判定

水谷らは，教科書コーパス語彙表を単語難易度の基準データとし，これを学習データとした機械学習によって単語の難易度を 20 段階に分類した [5]．教科書コーパス語彙表とは，教科書コーパスから抽出した語彙リストであり，小学校前半，小学校後半，中学校，高校の 4 つのカテゴリのテキストセットに出現する単語とその出現頻度が収録されている．機械学習には Support Vector Regression(SVR) が使用され，素性としてウェブ出現頻度，文書難易度初出・平均，漢検難易度初出の 3 種類が用いられた．これらは単語の難易度と高い相関関係があると水谷らは考えている．まず，単語の出現頻度が高くなるほど，単語の難易度が低くなるため，100 億文のウェブコーパスにおける単語の出現頻度を「ウェブ出現頻度」の素性とした．次に，文書の難易度が高くなるほど，そこに出現する単語の難易度も高くなる傾向があるため，文書難易度初出・平均を素性として用いた．文書難易度初出とは，コーパスを用いて各文書の難易度を推定したあと，単語が出現した文書のうち最も低い文書難易度と定義される．そして，文書難易度平均とは，その単語が出現する文書の文書難易度の重み付き平均値である．ただし，重みは文書における単語の出現頻度としている．最後に，ある単語が出現する日本漢字能力検定の問題文のうち最も簡単な級を「漢検難易度初出」の素性として用いた．実

験の結果から，文書難易度初出と漢検難易度初出が素性として有効に働くことが分かった．

2.1.2 外国人を対象とした日本語単語の難易度判定

ここでは，外国人の日本語学習者を対象とした日本語単語の難易度を策定する研究，日本語単語の難易度を利用してテキストの難易度を判定する研究，日本語単語の難易度そのものを判定する研究について紹介する．

Sunakawa らは，17,920 語の日本語教育用の日本語単語を収録した日本語教育語彙表を構築した [9]．作成時の基礎資料として，「現代日本語書き言葉均衡コーパス」と「日本語教科書コーパス」を用いた．前者は現代日本語の書き言葉の全体像を把握するために構築されたコーパスであり，書籍全般，雑誌全般，新聞，白書，ブログ，ネット掲示板，教科書，法律などの多様なジャンルのテキストから構成される 1 億 430 万語のコーパスである．後者は市販されている初級から上級までの教科書 100 冊の電子データ版から構成されるコーパスである．日本語教育語彙表に記載されている情報の一つとして，語彙の難易度がある．日本語教育上の語彙レベルが示され，初級前半，初級後半，中級前半，中級後半，上級前半，上級後半の 6 段階がある．

劉と内田は，日本語を学習する外国人を対象とした日本語テキスト難易度推定手法を提案した [8]．テキストの難易度は単語の難易度と 2 つの文節の係り受け距離を用いて推定する．これは，難しい単語が出現する，または文の構造が複雑になると，テキストの難易度も高くなるという仮定に基づいている．単語の難易度の尺度としては，テキストにおける日本語能力試験の各受験級 (1 級，2 級，3 級，4 級) を用いた．

中西らは VOD 講義で使われているスライドと発話の書き起こしテキストに出現する漢字 (文字) や単語に対して難易度を判定し，これを基に VOD 講義の難易度を推定する手法を提案した [6]．日本語単語の難易度は日本語能力試験における 4 つの試験区分と定義した．日本語能力試験の階級は，現在は N1 から N5 の 5 段階に分かれているが，ここでは旧試験の階級である 1 級から 4 級を試験区分としている．漢字や単語の試験区分が定義されている場合にはそれをそのまま利用し，定義されていないときは，日本語の辞書データやウェブ上の文を利用して学習されたサポートベクタマシン (Support Vector Machine;SVM) によって難易度判定を行う．SVM の学習パラメータ (学習素性) は共起語級別比，係り受け級別比，漢字単語難易度，漢字意味難易度の 4 つである．

共起語級別 まず，国語辞典における辞書定義文からなるコーパスを用意する．ある単語 w について，その単語と定義文内で共起する他の単語の試験区分 L の級ごとの比率を「共起語級別」とする．

係り受け級別比 上と同様に国語辞典における辞書定義文からなるコーパスを用いる．ある単語 w について，その単語と定義文内で直接の係り受け関係にある他の単語の試験区分 L の級の比率を「係り受け級別比」とする．

漢字単語難易度 単語 w を構成する漢字の試験区分のうち，最も難易度の高い試験区分を「漢字単語難易度」とする．

漢字意味難易度 単語 w の辞書の定義文に出現する漢字の試験区分の相乗平均を「漢字意味難易度」とする．辞書に登録されていない単語については，Google 検索でその単語を検索し，検索順位が最上位のウェブページからその単語を含む文を取得し，これをその単語の定義文とみなし算出する

SVM による難易度推定の性能を評価するために，試験区分データ (試験区分が既に分かっている単語) の再評価と未区分単語 (試験区分が分かっていない単語) に対する難易度推定を評価した．まず，単語ベースで得られた共起語級別比と係り受け級別比を用いて試験区分データの難易度を判定した場合，1 級と 4 級に判定されたものが多く，未区分単語に対してもいずれかに振り分けてしまう傾向があった．次に，文字ベースである漢字単語難易度を利用して試験区分データの難易度推定を行った場合，2 級に集まる傾向が強く，未区分単語でも 2 級が多いことが分かった．そして，漢字単語難易度と漢字意味難易度の両方を用いた場合，2 級に判定される傾向がありながら，試験区分データで 4 級のものは 4 級に正しく判定されるようになった．最後に，4 つの学習パラメータを全て利用した場合，単語ベースの学習パラメータを用いた手法で見られた 1 級と 4 級に判定が偏る傾向は，文字ベースの学習パラメータを組み合わせることである程度緩和された．

2.2 漢字処理における母語の影響

玉岡は，中国語と英語を母語とする日本語学習者を対象として，漢字二字熟語と外来語の処理の効率性を測定した [10]．実験では，中国語を母語とする日本語学習者，英語を母語とする日本語学習者および日本人大学生に対して，漢字二字熟語を一つずつスクリーンに提示し，それが正しい日本語の単語であるかどうかを「はい」または「いいえ」のキーを押すことで答えてもらった．各単語が提示されてからキーを押すまでの時間，ならびに判定の正誤を記録した．その結果，漢字で表記した条件での漢字二字熟語の処理速度について，中国語を母語とする日本語学習者は英語を母語とする学習者と比較して 826 ミリ秒も判断が速く，また判定の正解率も 7.6 ポイント高いことが分かった．この結果，中国語が母語である学習者は漢字表記の日本語を処理する能力が高いことが示された．しかしながら，同じ漢字二字熟語の単語を平仮名表記で提示した場合，処理時間には母語によって有意差が見られなかった．したがって，学習者の母語が中国語である影響は漢字の処理に限られていることが分かった．中国語を母語とする学習者に対しては，

中国語と日本語の漢字の文字や意味の違いを示すことで，漢字や漢字で表記された単語を効果的に学ぶことができる可能性がある．

2.3 テキストの類似度に関する研究

2つのテキストの類似度を測ることは，自然言語処理における様々な場面で必要とされるため，盛んに研究が行われている．本研究においても，単語の難易度を推定する際に，テキスト間の類似度を測る必要がある．本節では，本研究で使用するテキストの類似度を測る手法を紹介する．

Kusner らは，文書間の距離 (非類似度) の計算手法として Word Mover's Distance(WMD) を提案した [4]．WMD は当時の最新技術である単語の分散表現と Earth Mover's Distance(EMD) と呼ばれる最適化問題を結合した手法である．単語の分散表現とは，単語の意味を表すベクトル表現で，大規模なコーパスから学習される．代表的な単語の分散表現は Skip-gram モデルであり，Word2Vec と呼ばれるツールで学習できる．単語の分散表現を用いると，ベクトル間の類似度を計算することにより，単語間の意味的な類似度が計算できる．WMD ではこの性質を利用して文書間の距離を求める．文書間の距離を文書 d から文書 d' に全ての単語を輸送する最適輸送コストとして見立てる．つまり，一方の文書の単語を別の文書の単語に変換したとき，その単語のペア間の距離の累積として計算される．単語の分散表現によって単語をベクトル空間へ埋め込むことができるので，単語のペア間の距離はユークリッド距離で計算できる．すなわち，語 i と語 j の距離は $c(i,j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ という式で定義する．$c(i,j)$ は文書 d の単語 i から文書 d' の単語 j への移動コストともみなせる．また，WMD では，正規化された BoW(nBOW) を用いて文書をベクトルで表現する．$\mathbf{d}_i$ は，正規化された文書ベクトル $\mathbf{d}$ の i 番目の要素であり，単語 i の出現回数が文書全体の全ての単語の出現回数の和に占める割合と定義する．これは単語 i が持つ総量とみなす．文書 d の単語 i から文書 d' の任意の単語に全体または部分的に移動する．2 つの文書の単語数が一致しない場合は，移動先の文書の複数の語に分配する．文書 d の単語 i から文書 d' の単語 j への移動量は $\mathbf{T}_{ij}$ とする．文書 d の単語 i から文書 d' の全ての単語への移動量 $\sum_{j=1}^{n} \mathbf{T}_{ij}$ は $\mathbf{d}_i$ と一致している．同様に，文書 d' の単語 j が文書 d の全ての単語からもらった量の総和 $\sum_{i=1}^{n} \mathbf{T}_{ij}$ も $\mathbf{d}'_j$ と一致している．このような条件の下，文書間の距離は式 (2.1) で算出される．これは文書 d から文書 d' に全ての単語を輸送するときに必要な最小累積コストである．

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i,j)$$

$$\text{subject to:} \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i \in \{1,...,n\} \tag{2.1}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j \in \{1,...,n\}$$

このような制約を考慮した移動の最小累積コストは，線形計画問題に帰着でき，EMD の特殊な場合に相当する．EMD での輸送量上限は，荷物の総量または倉庫の総容量のどちらか小さいほうであるのに対して，WMD では正規化されているので，両方の総量とも 1 である．WMD はハイパーパラメータがないので，チューニングする必要がなくて使いやすい．そして，単語の意味的な類似度を比較的正確に測ることのできる単語の分散表現の知識を組み込むことで，文書間の距離も正確に測ることができる．

2.4 本研究の特色

2.1.1 項で紹介した関連研究 [5] は教科書コーパス語彙表を単語難易度の基準データとし，機械学習によって単語の難易度を 20 段階に分類した．日本語単語の難易度を判定する研究ではあるが，日本語母語話者にとっての難しさ，すなわち単語が表す概念の難しさを対象としており，日本語学習者にとっての難易度を判定する手法ではない．これに対して，本研究は単語が表す概念の難しさではなく，中国語母語話者が日本語を学ぶ際に意味を推測するときの難しさを判定する点が異なる．

2.1.2 項で紹介した関連研究のうち，Sunakawa らの研究 [9] と劉らの研究 [8] は，外国人の母語話者を対象としているが，日本語単語の難易度を直接推定するものではない．また，中西らの手法 [6] は，外国人の母語話者を対象に日本語単語の難易度を推定しているという点では本研究と共通しているが，全ての外国人を対象とした汎用的なもので，学習者の母語の違いは考慮されていない．これに対して，本研究では中国語を母語とする日本語学習者を対象とした日本語単語の難易度を推定する点に特長がある．

2.2 節にて紹介した関連研究 [10] は，中国語を母語とする日本語学習者は漢字の処理に優れていることを明らかにした．既に述べたように，中国語を母語とする学習者に対しては，中国語と日本語の漢字の文字や意味の違いを示すことで，単語を効果的に学ぶことができると考えられる．本研究は，漢字を使用する中国語を母語とする学習者の特性を考慮し，中国語母語話者に特化した日本語単語の難易度推定手法を考案する．

第3章　提案手法

本章では，漢字表記の日本語単語に対し，中国語母語話者から見た難易度を自動推定する手法の詳細について述べる．3.1 節では，日本語単語難易度の定義を説明する．3.2 節では，提案手法の概要を示す．3.3 節では，日本語単語の難易度の推定手法について述べる．このうち，日本語単語と中国語単語の語義の対応付けは最も重要な処理であるので，3.4 節でその詳細を説明する．

3.1　日本語単語難易度の定義

中国語を母語とする日本語学習者を対象とした日本語単語の難易度を定義する．本研究における単語の難易度は単語を習得するときの難しさ，すなわち単語を最初に学ぶ際に意味を推測する難しさを表すものとする．

文化庁による漢語の分類 [12] を基に，日本語単語の難易度を表 3.1 のように定義する．難易度のクラスは，難易度が低い順に，大きく S, O, D, N の 4 つのクラスに分かれている．漢字で表記された日本語単語に対して，それと同じ単語が中国語にない場合は N とする．それ以外のとき，日本語単語と中国語単語の語義を比較し，両者の語義の類似性に応じて S, O, D と判定する．また，クラス S, O, D は，日本語単語と中国語単語の漢字表記が完全に一致しているときは X-1，異なるとき (例えば「議論」と「议论」) は X-2 のように細分化する．基本的には，日本語単語と同じ単語が中国語にあり，その意味が似ているほど，その単語を初めて学習するときでも日本語単語の意味を推測でき，したがって難易度が低くなるという考え方に基づいている．

S, O, D, N の例を図 3.1 に示す．まず,「支持」は日本語と中国語の両方において『支える』意味と『他人の意見に賛成する』という意味を持っているため，難易度クラスは S になる．次に,「対象」は日中において『目標となるもの』という意味があるが，中国語では他に『恋人，結婚相手』という意味があるため，難易度クラスは O となる．それから,「用心」は日本語では『注意』という意味を持つが，中国語では『下心』という異なる意味を持つので，難易度クラスは D となる．上記の 3 つの単語のうち,「支持」と「用心」は漢字表記が日本語と中国語で同じであるのに対し,「対象」は中国語では漢字表記が異なり,「対 象」であるので，難易度クラスは正確には S-1, O-2, D-1 となる．最後に,「挨拶」は同じ漢字語が中国語には存在しないため，難易度クラスは N となる．

表 3.1: 本研究における日本語単語の難易度の定義

クラス	説明	漢字表記
S-1	Same(日本語単語と中国語単語の語義が全て同じ)	同じ
S-2		異なる
O-1	Overlap(日本語単語と中国語単語の語義の一部が同じ)	同じ
O-2		異なる
D-1	Different(日本語単語と中国語単語の語義が全く異なる)	同じ
D-2		異なる
N	Nothing (日本語と同じ単語が中国語にない)	–

図 3.1: S, O, D, N の例

クラス O に該当するのは，図 3.1 の「対象」のように，中国語単語の意味の数が日本語単語の意味の数より多い場合だけではない．日本語単語の語義数が中国語単語の語義数より多い場合と日本語単語と中国語単語の一部に重なりがある (日本語にしかない意味，中国語にしかない意味の両方ともある) 場合もある．これらの例を図 3.2 に示す.「不幸」は日中において『幸福でない』という意味があるが，日本語では他に『みうちの者に死なれること』という意味がある.「調子」は日本語と中国語の両方において『音楽の節回しや，話し声の，音の高低のぐあい』という意味を持っているが，日本語では他に『物事が進んでゆく時の，進行のぐあい，または勢い』という意味が，中国語では他に『議論する時の論調，趣旨』という意味がある．

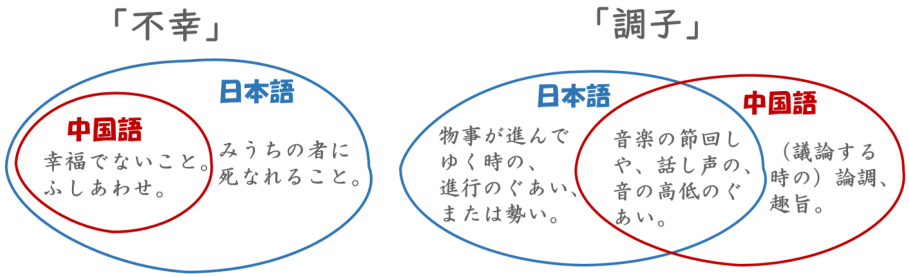


図 3.2: 他の O クラスの例

3.2 単語難易度の判定の概要

提案手法の概要を図 3.3 に示す．四角い枠で囲まれた項目は処理を表し，矢印の方向に進められる．まず，いくつかの日本語辞書と中国語辞書を用いて，辞書データを構築する．次に，漢字で表記されている日本語単語が与えられたとき，それと同じ表記を持つ中国語の単語を中国語辞書から検索する．日中で漢字の字体が異なるが実質は同じ表記である単語も検索する．中国語単語が見つかったとき，日中における語義の対応付けを行い，単語の難易度 (S, O, D) を推定する．それから，日本語と中国語で漢字表記が完全に一致しているかどうかによって難易度をさらに細分化する．最後に，単語の難易度を出力する．

なお，提案手法はプログラミング言語の Python 3.7.12 にて実装する．

3.3 単語難易度の推定

本節では単語難易度を推定する手法について述べる．3.3.1 項では辞書データの準備について述べる．3.3.2 項では単語難易度の推定のアルゴリズムについて説明する．

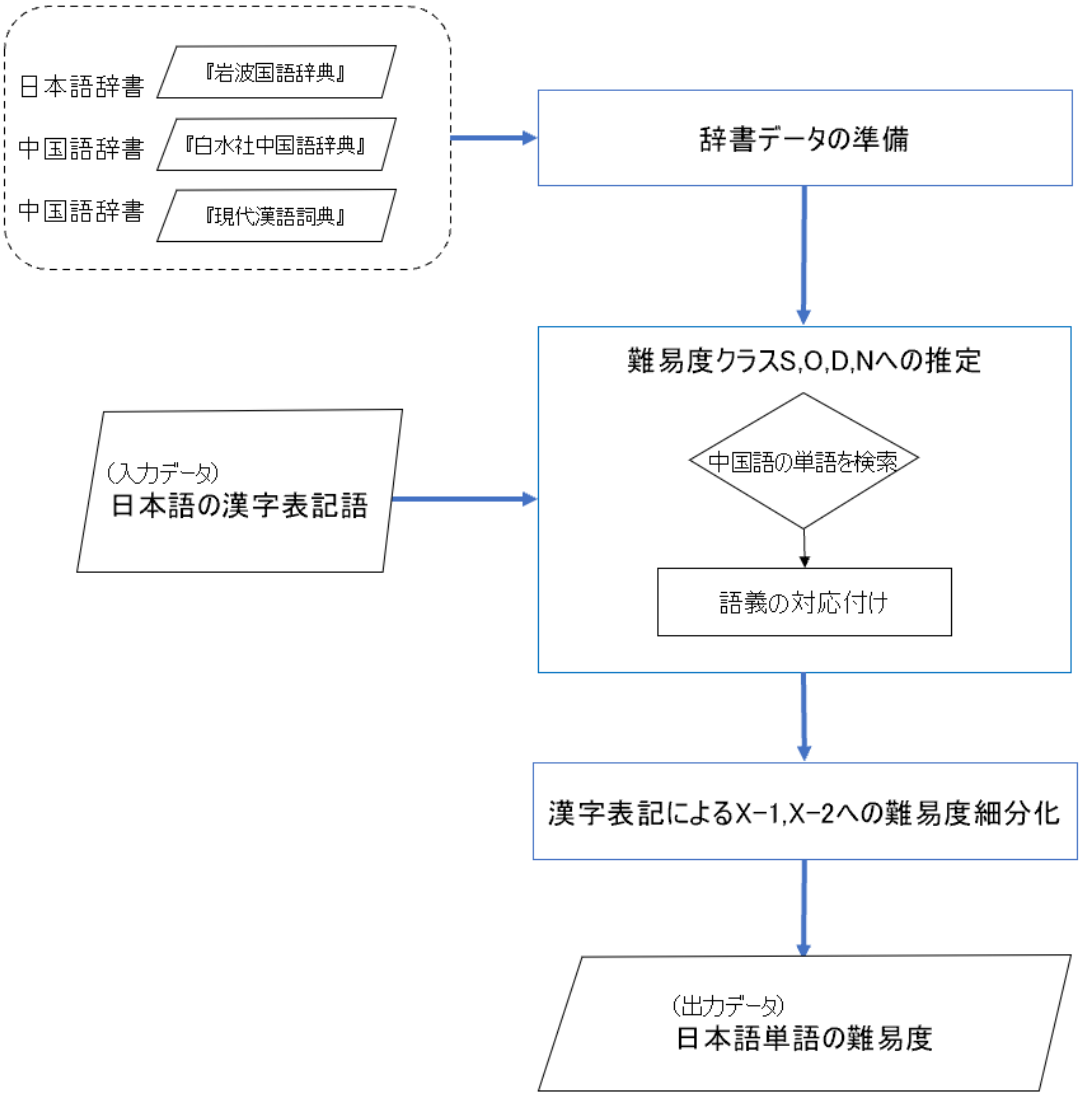

日本語辞書
『岩波国語辞典』
中国語辞書
『白水社中国語辞典』
中国語辞書
『現代漢語詞典』
辞書データの準備
難易度クラスS,O,D,Nへの推定
中国語の単語を検索
語義の対応付け
(入力データ)
日本語の漢字表記語
漢字表記によるX-1,X-2への難易度細分化
(出力データ)
日本語単語の難易度

図 3.3: 提案手法の概要

3.3.1 辞書データの準備

難易度の推定には日本語の語釈文と中国語の語釈文を必要とするため，日本語の辞書と中国語の辞書のデータを用意する．日本語の辞書として岩波国語辞典を，中国語の辞書として白水社中国語辞典[1]と現代漢語詞典 [7] を用いる．研究の初期の段階では白水社中国語辞典を中国語辞書として用いた．しかし，岩波国語辞典では語釈文が単語の意味を説明する文であるのに対して，白水社中国語辞典では中国語の単語に対して同じ漢字表記の日本語単語のみで構成される語釈文が多い．例えば中国語単語「医学」の語釈文は『医学.』である．そこで，中国語単語に関しても同じく意味の説明文の語釈文が書かれた辞書として，現代漢語詞典を採用した．各辞書の特徴を以下に示す．

- **岩波国語辞典**
  約 5 万 6 千の見出し語から構成される．それぞれの見出し語には，entry ID(見出し語の ID)，読み，品詞，語義の情報がある．語義の情報として，語義の ID と定義文（意味の説明文）が収録されている．

- **白水社中国語辞典**
  約 6 万 5 千の中国語について，意味や品詞，発音方法，用例などを調べることができる．語釈文は日本語で書かれている．Weblio 日中中日辞典に収録されている辞書の一つである．ただし，見出し語は中国語の簡体字が用いられているため，検索キーワードが日本語の漢字の場合，中国語の単語がヒットしないことがある．そのため，中国語の簡体字に変換した検索キーワードを用いるべきである．

- **現代漢語詞典**
  権威ある標準中国語の規範的な辞典である．辞書全体で約 65000 の中国語が収録されている．語釈文は中国語で書かれている．日本語単語と中国語単語の語義の対応付けを行うためには，語釈文はともに日本語で書かれている必要がある．そのため，現代漢語詞典の語釈文を Baidu 通用翻訳 API[2]を用いて日本語に翻訳する．また，見出し語が簡体字のため，日本語の単語を検索するには漢字の変換が必要である．

岩波国語辞典のデータから抽出した単語や語釈文などの例を表 3.2 に示す．語義の ID は 5 つの数字から構成されており，最初の 2 つは entry ID に対応する．残りの 3 つは語義の階層構造を表し，上位の桁ほど粗い粒度の語義を表す．語義 ID の末尾の数字が 0 である語釈文は 26192-0-0-2-0〔哲学〕のような分野を表すものもあれば，26353-0-0-1-0「標準の高さ」のような粗い粒度の語義もある．分野は語義ではないことが自明である．また，粗い粒度の語義に関しては，その下位の階

[1] https://cjjc.weblio.jp/category/cgkgj
[2] https://fanyi-api.baidu.com/

表 3.2: 岩波国語辞典の抜粋

entry id	見出し語	読み	品詞	語義	
				語義 ID	語義
26192 - 0	真理	しんり	名	26192-0-0-1-0	本当 の 事 。 間違い で ない 道理 。 正当 な 知識 内容 。
				26192-0-0-2-0	〔 哲学 〕
				26192-0-0-2-1	判断 内容 が もつ 客観 妥当 性 。 意味 の ある 命題 が 事実 に 合う こと 。
				26192-0-0-2-2	論理 の 法則 に かなっ て いる という 、 形式 的 な 正し さ 。
26353 - 0	水準	すいじゅん	名	26353-0-0-1-0	標準 の 高 さ 。
				26353-0-0-1-1	物事 の 価値 や 働き など を 調べる 時 の 基準 と なる 程度 。
				26353-0-0-1-2	世間 で 通用 し て いる 標準 。 「 生活 <EX> 水準 </EX> が 高く なる 」
				26353-0-0-2-0	高 さ に関する 位置 。
				26353-0-0-3-0	「 水準 器 」 の 略 。

層に位置する細かい粒度の語釈文がある．そのため，語義 ID の末尾が 0 の語義は，それより下位の語義がある場合には削除する．

白水社中国語辞典の語釈文の例を図 3.4 に示す．数字は語義の番号 (ID) を表す．「1」のような普通の数字の番号と①のような丸囲み数字の番号が同時に存在する場合，丸囲み数字の後ろでは語義の細かい使用場面 (用例) が示され，普通の数字の後ろに語釈文が書かれている．このような語釈文に関しては，前処理の段階では丸囲み数字の後ろの内容と普通の数字の後ろの語義と合併し，１つの語義になるように処理する．例えば，図 3.4 では丸囲み数字の後ろに書かれている「(空間の範囲を示す．)」,「(年齢・序列・程度・割合などの高さの範囲を示す．)」,「(数量の範囲を示す．)」と普通の数字の後ろの語義「... より上，... 以上.」と合併し，一つの語義とみなす．

現代漢語詞典はプレインテキストのデータである．ルールベースで品詞と語釈文を語義毎に抽出する．

岩波国語辞典，白水社中国語辞典，現代漢語詞典のいずれも，語釈文の中に単語の意味の説明以外の記述が含まれることがある．語釈文の類似度を測るとき，このような記述を残すのは望ましくない．そこで，語釈文から意味の説明ではない記述を除去するルールを作成した．そのルールの一部を表 3.3 に示す．また，全てのルールを付録 A に示す．これらのルールはパターンであり，これにマッチした文字列を削除することを意味する．

ルールならびにその使用例を図 3.5 に示す．岩波国語辞典の「電気」の語釈文のうち,「例えば」で始まる 2 番目の文は電気の意味を説明するための例で，意味の説明ではない．これは「例えば、～。」というルールで削除される．また，白水社中国語辞典の「一時」の語釈文のうち,（と）で囲まれた記述は文法に関する注釈であり，意味の説明ではない．これは「(～用い)」というルールで削除される．

3.3.2 単語難易度の推定

日本語単語難易度を推定する処理の流れを以下に示す．

1. 難易度を推定する日本語単語を w^J とする．ここでは w^J は漢字で表記された単語に限る．

2. 中国語辞書を検索し，w^J と同じ表記を持つ中国語の単語 w^C を検索する．もし見つからない場合，日中漢字マッピングテーブル [3] を参照し，日本語の漢字を中国語の漢字に変換してから，w^C を検索する．中国語の漢字には簡体字と繁体字があるが，本研究で用いる中国語辞書では見出し語は簡体字で表記されているため，ここでは簡体字に変換する．

3. w^C が見つからない場合，w^J の難易度クラスを N と判定する．それ以外は次のステップへ．

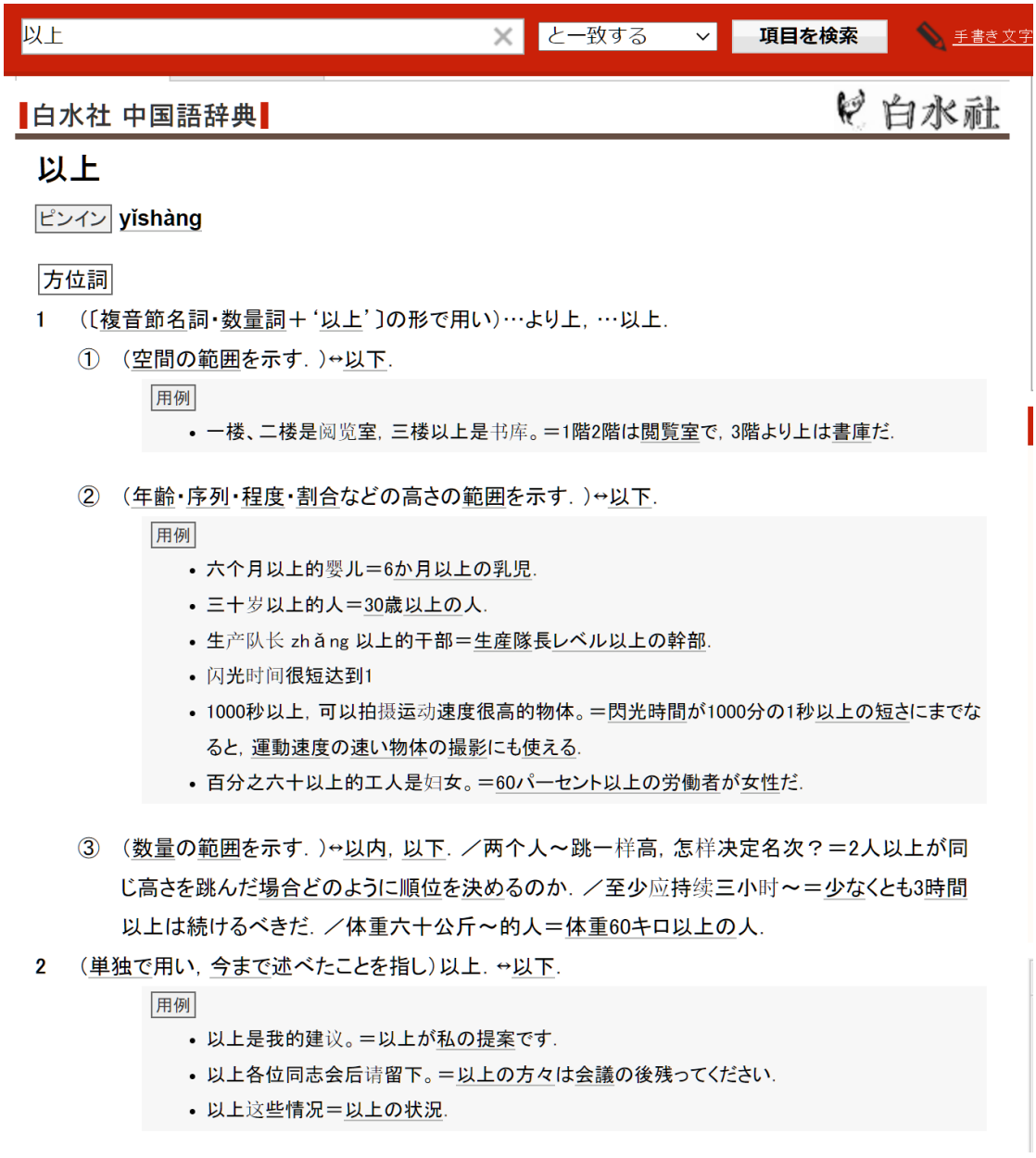


図 3.4: 白水社中国語辞典の語釈文の例

表 3.3: 語釈文以外の情報を削除するルールの抜粋

辞書	タイプ	ルール
岩波国語辞典	例	例えば、〜。 例、〜。 （例えば、〜） （例、〜）
	読み	「〜」と読む。 「〜」とよむ。
白水社中国語辞典	用法	（〜用いる.） （〜用いて） （〜用い） （〜用いるが, 〔〜の形で用いる； 〔〜の形で用い, ‘〜’の形で用いる.
	読み	（方言では〜）
現代漢語詞典	用法	（〜的） （〜地） （〜儿） （〜儿的）
	呼び方	所以叫〜。 也叫〜。

岩波国語辞典

ルール:	例えば、〜。
見出し語:	電気
元の語釈文:	広く電荷の関連する現象、およびその現象の実体。例えば絹で摩擦したガラス棒が紙片を引きつけるような現象を起こさせる原因となるもの。エネルギーの一つの形態。
修正された語釈文:	広く電荷の関連する現象、およびその現象の実体。エネルギーの一つの形態。

白水社中国語辞典

ルール:	（〜用い）
見出し語:	一時
元の語釈文:	（〔複音節名詞・数量詞＋‘以上’〕の形で用い）…より上，…以上.
修正された語釈文:	…より上，…以上.

図 3.5: ルールによって説明以外の記述を除去する例

4. 日本語の辞書から w^J の語釈文の集合 $S^J = \{s_1^J, \cdots, s_n^J\}$ を得る．s_i^J は w^J の i 番目の語義の語釈文を表す．同様に，中国語の辞書から w^C の語釈文の集合 $S^C = \{s_1^C, \cdots, s_m^C\}$ を得る．

5. 日中それぞれの語義の集合 S^J と S^C に対して，語義の対応付けを行い，その結果に応じて w^J を S, O, D のいずれかに分類する．この処理の詳細は 3.4 節で詳しく述べる．

6. w^J と w^C の表記が一致しているときは S-1, O-1, D-1 のいずれかを，一致していないときは S-2, O-2, D-2 のいずれかを難易度クラスとする．

以上の手続きをフローチャートとしてまとめたものを図 3.6 に示す．

3.4 語義の対応付け

本節では，日本語単語の語義の集合 S^J と中国語単語の語義の集合 S^C が与えられたとき，両者の語義の対応付けを行う手法について述べる．3.4.1 項では，2 つの語義の類似度を計算する方法について述べる．3.4.2 項では，語義の対応付けのアルゴリズムを説明する．3.4.3 項では，品詞を考慮した語義の対応付けの手法を紹介する．3.4.4 項では，多対多を考慮した語義の対応付けの手法を述べる．

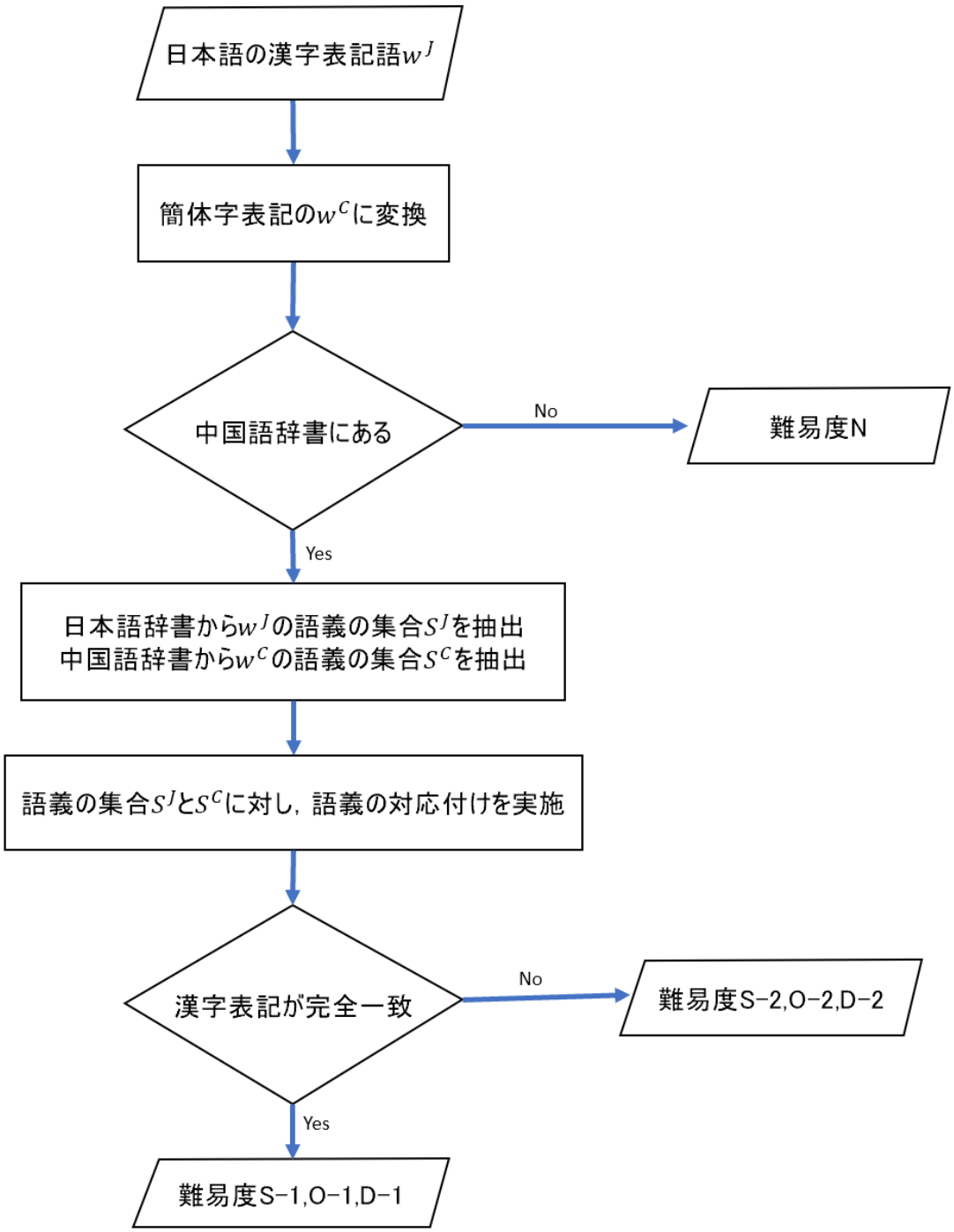


図 3.6: 単語難易度の推定のフローチャート

3.4.1 語義の類似度

語義の対応付けを行う際，日本語単語の語義と中国語単語の語義の類似度を計算する．本研究では，語義のベクトル間類似度を用いる手法と，Word Mover's Distance(WMD) を用いる手法の 2 つを採用する．

まず，語義のベクトル間類似度を用いる手法について説明する．日本語単語および中国語単語の語義に対し，それをベクトルで表現する．以下，これを語義ベクトルと呼ぶ．語義ベクトルは，その語義の語釈文に含まれる単語の分散表現の平均ベクトルで計算する．単語の分散表現は，大量のコーパスから学習された単語の意味を表すベクトル表現である．本研究では，単語の分散表現として NWJC2vec [1] を利用する．『国語研日本語ウェブコーパス』(NWJC) [2] はウェブを母集団として構築した 258 億語規模の日本語コーパスである．NWJC2vec は，NWJC から Word2Vec および fastText によって学習された単語埋め込みデータである．公開されているデータは，200 次元 CBOW (word2vec), 200 次元 skip-gram (fastText), 300 次元 CBOW (fastText), 300 次元 skip-gram (fastText) からなる．本研究では 200 次元 CBOW (word2vec) を用いる．語釈文 s^J から得られた語義ベクトルと語釈文 s^C から得られた語義ベクトルをそれぞれ $\vec{v}^J, \vec{v}^C$ とし，両者のコサイン類似度を語義間の類似度とする．$\vec{v}^J$ と $\vec{v}^C$ のコサイン類似度は式 (3.1) によって定義される．

$$cos\theta_{\vec{v}^J,\vec{v}^C} = \frac{\vec{v}^J \cdot \vec{v}^C}{|\vec{v}^J||\vec{v}^C|} \tag{3.1}$$

ここで θ は 2 つのベクトルの間の角度を表，θ が小さいほど (2 つのベクトルの方向が近いほど)$cos\theta$ は大きな値を取る．コサイン類似度は 2 つのベクトルの類似度を求めるときによく使われる尺度で，-1 から 1 までの値を取り，1 に近ければ近いほど類似度が高い．

次に，WMD を用いる手法を説明する．WMD は 2.3 節で紹介した関連研究 [4] の手法であり，文書間の距離，すなわち非類似度を測ることができる．ここでは 2 つの語義の語釈文を文書とみなし，語釈文間の距離を語義間の非類似度とみなす．WMD では，文書 s に異なる n 個の単語がある場合，文書 s の各単語 $w_1, \cdots, w_n$ には $\frac{1}{n}$ の重みがある．文書 s の単語 w_i から文書 s' の単語 w'_j への輸送コストは単語ベクトル間のユークリッド距離である．ここでは，語義のベクトル間類似度を用いる手法と同じく，単語ベクトルとして NWJC2vec を用いる．そして，文 s から文 s' に全ての単語を輸送する最小の輸送コストが WMD によって計算される文書間の距離となる．

3.4.2 語義の対応付け

日本語単語の語義の集合 S^J と中国語単語の語義の集合 S^C から難易度クラス S, O, D のいずれかを判定するアルゴリズムを Algorithm 1 に示す．

Algorithm 1 S, O, D の判定

procedure DIFFICULTY-CLASS(S^J,S^C)
  $SA = \emptyset$
  while $S^J \neq \emptyset$ and $S^C \neq \emptyset$ **do**
    $(i', j') \leftarrow \arg\max_{(i,j)\ s.t.\ s_i^J \in S^J, s_j^C \in S^C}\ sim(s_i^J, s_j^C)$
    $sim_{sense} \leftarrow sim(s_{i'}^J, s_{j'}^C)$
    if $sim_{sense} \geq T_m$ **then**
      $SA \leftarrow SA \cup \{\ (s_{i'}^J, s_{j'}^C, sim_{sense})\ \}$
      $S^J \leftarrow S^J \setminus \{s_{i'}^J\}$
      $S^C \leftarrow S^C \setminus \{s_{j'}^C\}$
    else
      break the loop
    end if
  end while
  if SA is $\emptyset$ **then**
    return D
  else if $S^J \neq \emptyset$ or $S^C \neq \emptyset$ **then**
    return O
  else
    return S
  end if
end procedure

4 行目の $sim(s_i^J, s_j^C)$ は 2 つの語釈文の類似度を表す．日本語の語義と中国語の語義の全ての組み合わせのうち，$sim(s_i^J, s_j^C)$ が最も高くなる語義の組を (i', j') とし (4 行目)，その最大の語義間の類似度を sim_{sense} とする (5 行目)．日本語の i' 番目の語義と中国語の j' 番目の語義は同じ意味を持つとみなして対応付け，集合 SA に追加する (7 行目)．SA は対応付けされた語義の組とその組の語義間類似度を記録する．対応付けられた語義を S^J と S^C から除き (8,9 行目)，残りの語義について同じ処理を繰り返す．ただし，sim_{sense} が閾値 T_m 以上ではないとき，つまり語義間の類似度が十分に大きくないとき，2 つの語義は同じ意味を持つとはみなさず，語義の対応付けを終了する (6,11 行目)．閾値 T_m は実験的に決定する．詳細は 4 章で述べる．

語義の対応付けの終了後，14〜20 行目では，対応付けできる語義の組が 1 つも見つからないときは D と判定し，対応付けできる語義の組はあるが全ての語義について対応付けできないときは O と判定し，全ての語義について対応付けができたときは S と判定する．

Algorithm 1 は語義の類似度をベクトル間類似度で計算することを仮定しているが，語義の類似度を WMD で測るときも同様の処理を行う．ただし，WMD で算出されるのは距離 (非類似度) なので，類似度を非類似度に置き換える修正が必要である．4 行目では，距離が一番小さい語義の組 (i', j') を選択する．6 行目では，距離が閾値 T_m 以下であるときに 7〜9 行目の処理を行う．すなわち，距離が閾値 T_m より大きければ，語義の対応付けを行わない．

語義の対応付けの例を図 3.7 に示す．この例では単語「圧力」について語義の対応付けを行っている．日本語単語の「圧力」には 2 つの語義が，中国語単語「圧力」には 3 つの語義がある．語義間の線上の数値は語義のベクトル間類似度で計算された語義の類似度である．類似度が最も大きい語義の組は s_2^J と s_2^C なので，まずこれらを対応付ける．残された語義の中で類似度が最も大きい語義の組は s_1^J と s_1^C なので，次にこれらを対応付ける．日本語の語義はもう残されていないため，対応付けの処理を終了し，難易度クラスを O と判定する．なお，閾値 T_m を 0.2 に設定しているとき，類似度がそれ以下の語義の組 (この例では s_1^J と s_3^C) は必ず対応付けないことにする．

Algorithm 1 は，中国語の辞書として白水社中国語辞典と現代漢語詞典を用いたときのそれぞれについて適用する．2 つの辞書による判定結果が異なる場合は，式 (3.2) に示す語義の対応付けのスコア (対応付けられた語義間の類似度の平均値) を算出し，大きい方の判定結果を採用する．

$$\text{sense-align-score} = \frac{1}{|SA|} \sum_{(*,*,sim_k) \in SA} sim_k \tag{3.2}$$

2 つの中国語辞書を用いたときの判定結果が異なるときに最終的な難易度クラスを決定する例を図 3.8 に示す.「趣味」という単語に対して，岩波国語辞典と白水社中国語辞典の語義の対応付けを行った結果，S クラスと判定した．一方で，岩波

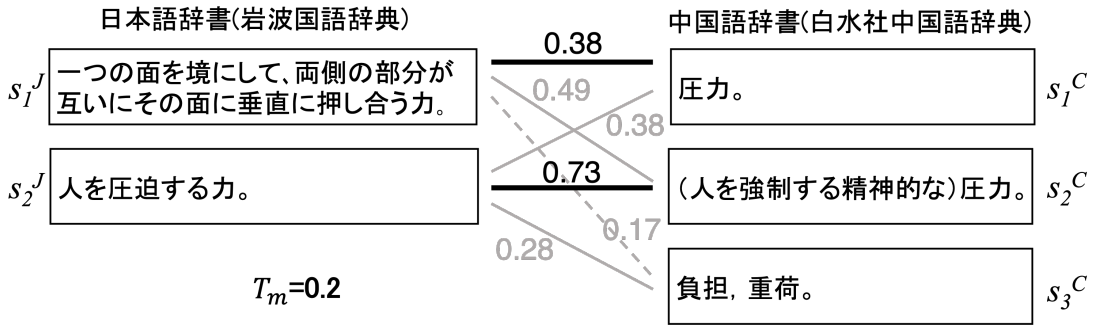


図 3.7: 「圧力」の語義の対応付けの例

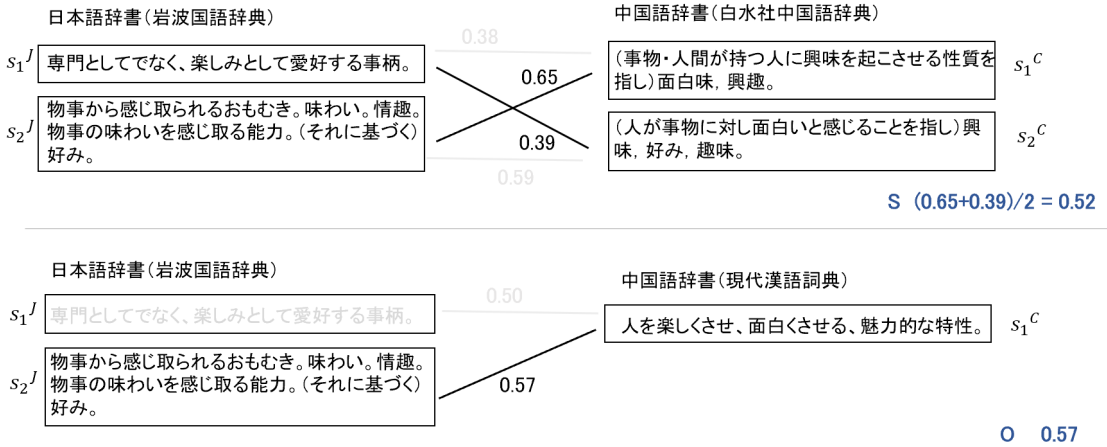


図 3.8: 「趣味」の語義の対応付けの例

国語辞典と現代漢語詞典による判定結果は O クラスであった．式 (3.2) に示す語義の対応付けのスコアを算出すると，白水社中国語辞典の場合は 0.52，現代漢語詞典では 0.57 となる．後者の方がスコアが高いので,「趣味」の難易度クラスは O と判定する．

3.4.3 品詞を考慮した語義の対応付け

岩波国語辞典では，1 つの見出し語が複数の品詞の語義を持つことが散見される．それに対して，中国語の辞書では，見出し語に対して 1 つの品詞の語釈文だけが記載されることもあれば，品詞毎に分けられて語釈文が載せられることもある．表 3.4 にその例を示す．日本語単語の「理解」や「意識」は，いわゆるサ変名詞あるいはサ変動詞で，名詞として働くこともあれば動詞として働くこともあるが，岩波国語辞典では両者は区別されずに語義が定義されている．一方，中国語の「理解」は日本語のような名詞としての意味がありながらも，中国語辞書では動詞としての語釈文しかない．その一方で,「意識」では名詞と動詞両方の語釈文が中国語辞書に載せられている．

表 3.4: 多品詞語の語釈文の例

単語	岩波国語辞典		白水社中国語辞典	
	品詞	語釈文	品詞	語釈文
理解	名・ス他	物事のすじみちをさとること．わけを知ること．物事がわかること．	動詞	(多く書き言葉に用い；人間・事物の状況をよく考えて) 理解する，わかる．
		人の気持や立場がよくわかること．		
意識	名・ス他	自分が現在何をやっているか、今はどんな状況なのかなどが自分でわかる、心の働き。また、その働きで自分にわからせること。	名詞	((哲学))(認識・判断などの精神の働きを指し) 意識．
				(大脳が目覚めて精神が活動している状態を指し) 意識．
			動詞	気づく，はっきりと知る．

これまでの手法では，語義の対応付けを行う際に品詞を考慮していなかった．すなわち，表 3.4 の例では，中国語の「意識」は 3 つの語義を持つ単語として扱っていた．しかし，異なる品詞の語釈文は，本来は対応付けるべきではないのに誤って対応付けされる可能性がある．また，中国語単語の名詞の語義と動詞の語義が似ている場合，品詞を考慮しないと 1 つの単語が似ている語義を複数持つことになるが，このような状態で日本語単語の語義との対応付けを行うと，中国語の似ている語義のうちどちらか一方しか日本語単語の語義に対応付けられないため，S, O, D の判定を誤る可能性が高くなる．

そこで，品詞を考慮した語義の対応付けの手法を提案する．以下にその手続きを示す．

1. 日本語単語と中国語単語において一致する品詞がない場合，3.4.2 項で紹介した Algorithm 1 にしたがって，品詞を考慮せずに難易度を判定する．

2. 一致する品詞がある場合，品詞ごとに Algorithm 1 を適用し，難易度クラスを判定する．その後，式 (3.2) に示す語義の対応付けのスコアを算出する．難易度クラスと語義の対応付けのスコアを合わせて 1 つの判定結果とする．

3. 中国語の辞書として白水社中国語辞典と現代漢語詞典を用いたとき，1 つの品詞について，2 つの判定結果がある．2 つの辞書による判定結果が異なる

場合，語義の対応付けのスコアが大きい結果を採用し，1 つの品詞に対応する判定結果が 1 つになるようにする．

4. 複数の品詞がある場合，品詞の数だけ判定結果がある．その中から語義の対応付けのスコアが最も大きい難易度クラスを選択する．ただし，この段階で S と D が同時に存在する場合は，語義の対応付けのスコアと関係なく，O と判定する．その理由を以下に述べる．難易度クラスが D と判定されたとき，対応する語義がひとつもないため，語義の対応付けのスコアは 0 となる．したがって，語義の対応付けのスコアの大きい難易度クラスを選択すると，常に S が選ばれることになる．ところが，このような状況下では，ある品詞 A の語義は全て対応付けられている (クラス S のとき) が，別の品詞 B の語義は対応付けられていない (クラス N のとき)．単語全体で見れば，一部の語義のみが対応付けられている状態であり，これは O に該当する．よって，品詞毎の判定が S と D に分かれたときは，最終判定を O にするというルールを設定する．

3 つの辞書において，品詞の表記は異なることに注意を要する．その例は前述の表 3.4 でも確認できる．名詞は，岩波国語辞典では「名」，白水社中国語辞典では「名詞」と表記されている．また，動詞は，国語辞典では「ス他」，白水社中国語辞典では「動詞」と表記されている．「ス他」は，サ変動詞であることと他動詞であることを表している．品詞を考慮した語義の対応付けの手法では，語義を品詞別に処理することが必要なので，3 つの辞書データにおける品詞の表記を統一しなければならない．そのために，品詞対応表を作成し，それに基づいて表記を統一した．本研究で作成した品詞対応表を表 3.5 に示す．例えば中国語の辞書に「量詞」と「量」があるが，岩波国語辞典では「名」と表記されている．これらは全て「名詞」という表記に変換した．

3.4.4 多対多を考慮した語義の対応付け

前述の提案手法では語義の対応関係に関して，1 対 1 の対応のみを認めている．しかし，語義の対応関係は必ずしも 1 対 1 ではなく，1 対多や多対多になることがある．これは，本研究で採用した日本語辞書と中国語辞書において，語義の粒度が異なるからである．つまり，単語の意味に対して，ある辞書では 1 つの語釈文のみで表すが，ほかの辞書ではそれを分けて 2 つ以上の語釈文で表すことがある．前節に示した表 3.4 における「理解」という単語がその一例である．「理解」の語義の正しい対応付けを図 3.9 に示す．岩波国語辞典の語釈文 s_1^J と s_2^J は，白水社中国語辞典の語釈文 s_1^C と対応付けるべきである．なぜなら，語釈文 s_1^C は語釈文 s_1^J と s_2^J をまとめた語義の定義とみなせるからである．

このように，語義の対応関係は必ずしも 1 対 1 ではなく，1 対多や多対多になる可能性もある．語義の粒度の相違に関して，2 つの中国語辞書を用いることで問題

表 3.5: 品詞対応表

品詞	岩波国語辞典	白水社中国語辞典	現代漢語詞典
名詞	名/名 (038)	名詞	名
			名时间词
			数
		数詞＋量詞	数量词
		量詞	量
		付属形態素	名方位词
		方位詞	
動詞	ス自	動詞	动
	ス他		
	ス自他		
	ス他自		
形容詞	ダナ/ダ (047)	形容詞	形
			形状态词
			形属性词
副詞	副	副詞	副
	副 (038)		

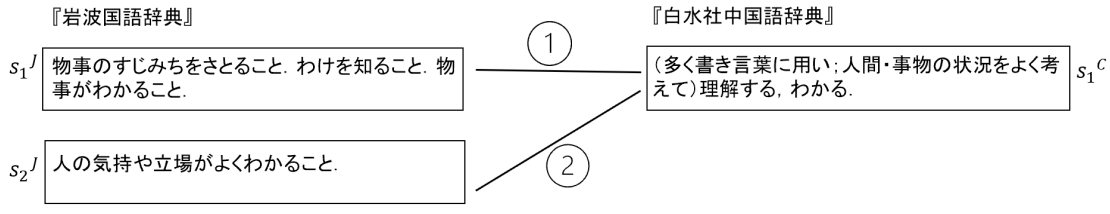


図 3.9: 「理解」の語義の対応付けの例

は緩和されうるが，語義の対応付けのとき，1 対多や多対多の対応付けもできれば，難易度クラスの判定の性能が向上することが期待できる．

多対多の対応を考慮したうえで，日本語単語の語義の集合 S^J と中国語単語の語義の集合 S^C から難易度クラス S, O, D のいずれかを判定するアルゴリズムを Algorithm 2 に示す．

Algorithm 2 多対多の対応付けを考慮した S, O, D の判定

procedure DIFFICULTY-CLASS2(S^J,S^C)
  $\mathbf{L}^J = \{0, .., i, ..., |S^J| - 1\}$
  $\mathbf{L}^C = \{0, .., j, ..., |S^C| - 1\}$
  $\mathbf{A}_{pattern} \leftarrow \mathbf{L}^J \times \mathbf{L}^C = \{(i, j) | i \in \mathbf{L}^J, j \in \mathbf{L}^C\}$
  $ali(s_i^J, s_j^C) = 1$ or 0
  $\mathbf{A}_{align} \leftarrow$ array of all combination of$(a_1, ..., a_{|\mathbf{A}_{pattern}|})$ $s.t.$ $a_k = ali(s_i^J, s_j^C)$
  for $x = 0, \ldots, |\mathbf{A}_{align}| - 1$ **do**
    $score_x = 0$
    for $a_k \in \mathbf{A}_{align}[x]$ **do**
      if $a_k = 1$ **then**
        $score_{aligned} \leftarrow \{sim(s_i^J, s_j^C) | i, j \in \mathbf{A}_{pattern}[x]\}$
        $score_x \leftarrow score_x + score_{aligned}$
      else
        $score_{unaligned} \leftarrow \max\{0,\ 0.5 - sim(s_i^J, s_j^C) \mid i, j \in \mathbf{A}_{pattern}[x]\}$
        $score_x \leftarrow score_x + score_{unaligned}$
      end if
    end for
    $AlignmentScore[x] \leftarrow score_x$
  end for
  $x' \leftarrow \arg\max_{0 \le x \le |\mathbf{A}_{align}|} AlignmentScore[x]$
  if $a_k = 1$ for all k in $\mathbf{A}_{align}[x']$ **then**
    return S
  else if $a_k = 0$ for all k in $\mathbf{A}_{align}[x']$ **then**
    return D
  else
    return O
  end if
end procedure

まず，多対多の対応関係を考慮した全ての語義の対応付け（アライメント）の候補を生成する．2 行目の L^J は日本語単語の語釈文のインデックスの集合，3 行目 L^C は中国語単語の語釈文のインデックスの集合である．日本語の語義と中国語の語義の全ての組み合わせパターンを L^J と L^C にあるインデックスの組み合わせ

の集合 $\mathbf{A}_{pattern}$ で表す (4 行目)．5 行目の $ali(s_i^J, s_j^C)$ は対応関係の有無を表す数字の集合である．$ali(s_i^J, s_j^C) = 1$ は対応関係があるとし，$ali(s_i^J, s_j^C) = 0$ は対応関係がないとする．a_k を $ali(s_i^J, s_j^C) = 1$ または $ali(s_i^J, s_j^C) = 0$ とし，$\mathbf{A}_{pattern}$ におけるそれぞれの語義の組について対応関係があるどうかを $(a_1, ..., a_{|\mathbf{A}_{pattern}|})$ で表す．これが S^J と S^C に対する 1 つの語義のアライメントの候補を表す．全てのアライメントのパターンの集合 (a_k を 1 とするか 0 とするかの全ての組み合わせ) を $\mathbf{A}_{align}$ とする (6 行目)．

次に，個々のアライメントの候補のスコアを計算する．繰り返しになるが，a_k は $\mathbf{A}_{pattern}$ の k 番目の語義の組み合わせが対応付けられているかどうかを示す．k 番目の語義の組み合わせパターンは，i 番目の日本語単語の語義 s_i^J と j 番目の中国語単語語義 s_j^C とする．$a_k = 1$ のとき，s_i^J と s_j^C には対応関係があり，対応付けられたときのスコア $score_{aligned}$ を計算する (11 行目)．ここで $score_{aligned}$ は s_i^J と s_j^C の類似度 $sim(s_i^J, s_j^C)$ とする．s_i^J に含まれる単語の分散表現の平均ベクトルを $\vec{v}_i^J$，同様に s_j^C の文のベクトルを $\vec{v}_j^C$ とし，2 つのベクトルのコサイン類似度を $sim(s_i^J, s_j^C)$ とする．$a_k = 0$ のとき，s_i^J と s_j^C には対応関係がなく，対応付けないときのスコア $score_{unaligned}$ を計算する (14 行目)．$score_{unaligned}$ は 0 と $0.5 - sim(s_i^J, s_j^C)$ のうち，より大きい値とする．x 番目のアライメントのパターン $\mathbf{A}_{align}[x]$ に関して，対応付けられた語義の組のスコアと対応付けない語義の組のスコアを合わせて，スコアの和 $score_x$ を計算する (12,15 行目)．これを $\mathbf{A}_{align}[x]$ のスコア $AlignmentScore[x]$ とする (18 行目)．全てのアライメントの候補のうち，スコアが最大となる候補のインデックスを x' とおく (20 行目)．これは，$\mathbf{A}_{align}$ の x' 番目のアライメントのパターンのスコアが最大であることを意味する．$\mathbf{A}_{align}[x']$ に 0 と 1 があるときは，対応付けできる語義の組はあるが全ての語義について対応付けできているわけではないので，O と判定する．1 のみがあるときは全ての語義について対応付けができているため，S と判定する．0 のみがあるときは対応付けできる語義の組が 1 つも見つからなかったので，D と判定する．

コサイン類似度の最大値は 1 なので，$score_{unaligned}$ は $0.5 - sim(s_i^J, s_j^C)$ ではなく $1 - sim(s_i^J, s_j^C)$ と定義した方が自然である．4 章で後述するが，3.4.2 項における閾値 T_m は，実験ではおよそ 0.2 に設定する．これは，1 対 1 の対応のみを考慮した語義のアライメント手法では，語義ベクトルのコサイン類似度が 0.2 以上のときには語義を対応付けることを意味する．一方，$score_{unaligned}$ を $1 - sim(s_i^J, s_j^C)$ と定義すると，例えば $sim(s_i^J, s_j^C) = 0.3$ のとき，$score_{unaligned}$ は 0.7 となり，語義を対応付けないときのスコアが語義を対応付けるときのスコアよりも高くなる．これは不自然である．そのため，$sim(s_i^J, s_j^C)$ が 0.2 付近の値のとき (正確には 0.2 から 0.5 までの範囲にあるとき)，$score_{aligned}$ が $score_{unaligned}$ よりも大きくなるように，$score_{unaligned} = 0.5 - sim(s_i^J, s_j^C)$ と定義した．

多対多を考慮した「理解」の語義の対応付けを表 3.6 に示す．①と②はそれぞれ図 3.9 にある語釈文 s_1^J と s_1^C，s_2^J と s_1^C の対応関係を表す．①と②の語義の組で対応関係がある場合とない場合の全ての組み合わせは 4 通りとなる．対応関係

があるときのスコアが語義のベクトル間類似度で，それぞれ $sim(s_1^J, s_1^C) = 0.596$ と $sim(s_2^J, s_1^C) = 0.595$ となる．一方，対応関係がないときのスコアはそれぞれ $max\{0, 0.5 - sim(s_1^J, s_1^C)\} = 0$ と $max\{0, 0.5 - sim(s_2^J, s_1^C)\} = 0$ となる．$\mathbf{A}_{align}[0]$ から $\mathbf{A}_{align}[3]$ のうち，①と② の両方に対応関係があるとき，すなわち s_1^J と s_2^J を s_1^C と対応付けたとき，スコア $AlignmentScore$ が最大となった．全ての語義が対応付けられたため，難易度クラスを S と判定する．

表 3.6: 多対多を考慮した「理解」の語義の対応付け

	①$ali(s_1^J, s_1^C)$	score	②$ali(s_2^J, s_1^C)$	score	$AlingmentScore[x]$
$\mathbf{A}_{align}[0]$	1	0.596	1	0.595	1.191
$\mathbf{A}_{align}[1]$	1	0.596	0	0	0.596
$\mathbf{A}_{align}[2]$	0	0	1	0.595	0.595
$\mathbf{A}_{align}[3]$	0	0	0	0	0

第4章　評価

提案手法を 2 つの観点から評価する．ひとつは，提案手法によって日本語単語の難易度をどの程度正確に判定できるかという観点である．4.1 節では，提案手法による難易度判定の正解率を測ることによって提案手法の性能を評価する実験について報告する．もうひとつは，提案手法で設計した日本語単語の難易度が実際の学習者から見てどれだけ妥当であるかといった観点である．4.2 節では，中国語を母語とする様々な習熟度レベルの日本語学習者を対象にアンケート調査を実施し，提案手法の単語難易度の尺度の妥当性を評価する．

4.1　難易度判定の評価

3 章では，語義の対応付けによる単語難易度の判定について説明した．本節では，提案手法による難易度クラスの判定の性能を評価する．N クラスとそれ以外の判定，ならびに S-1, O-1, D-1 と S-2, O-2, D-2 の判定は自明なので，ここでは難易度クラス S, O, D の判定のみを評価する．

4.1.1　データセットの作成

評価実験にあたり，文化庁による「中国語と対応する漢語」の資料 [12] を利用して，評価実験のためのテストデータを作成した．

「中国語と対応する漢語」は，初・中級の段階で学習者に教える漢語 1,882 語に対し，日中両言語における意味を比較して，これらを S, O, D, N のいずれかに分類している．その中から，S, O, D に分類される漢語をランダムに選択した．それぞれのクラスから 200 語を目安に抽出しようとしたが，O と D に分類された単語は 200 語未満であったため，その全てを抽出した．その結果，S クラスが 200 語，O クラスが 67 語，D クラスが 61 語，合計 328 語を抽出した．次に，この中から日本語辞書と中国語辞書に記載がない単語を除いた．すなわち，提案手法によって難易度クラスが N に分類される単語を除去した．また，文化庁による分類を見直したところ，辞書に掲載された語義のうち，あまり使われない語義は対応付けの対象としていない．その一方で，本研究では辞書に記載がある全ての語義を対応付けの対象とする．すなわち，文化庁によって定義された単語の意味関係の分類は提案手法の評価データとして適切ではない可能性がある．そのため，実験で使

用する日本語と中国語の辞書の語釈文を参照しながら，人手で S, O, D の分類をやり直した．この際に，語義の使用頻度に関わらず，辞書に記載された語義は全て対応付けの対象とした．最終的に 279 語の漢字表記語からなるデータセットを用意した．難易度クラスの内訳は，S クラスが 136 語，O が 121 語，D が 22 語となった．テストデータの一覧を表 4.1 に示す．

4.1.2 実験手順

4.1.1 項で説明したテストデータにおける評価単語について，提案手法によって難易度クラスを判定し，人手で付与した正解の難易度クラスと一致する割合を正解率とする．この正解率を評価基準として提案手法の性能を評価する．

単語の難易度を推定するための前処理として，語釈文の形態素解析が必要である．形態素解析とは，文を言語的に意味を持つ最小単位の形態素に分解し，それぞれの品詞や活用の変化などを判別することである．ここでは形態素はおおむね単語とみなせる．本実験で形態素解析を行う際には，一般的に使われている形態素解析器 MeCab-0.996.5[1]を採用した．MeCab は，辞書，コーパスに依存しないという意味で汎用的に設計されており，高い精度で形態素の解析ができる．MeCab を用いて形態素を解析することによって，語釈文を単語に分割し，それぞれの単語の品詞を同定する．この段階では，ストップワードを除去し，単語の基本形を抽出する．

3.4.2 項で述べた手法で日本語単語の語義と中国語単語の語義を対応付ける際，閾値 T_m を設定した．これは，語義のベクトルのコサイン類似度によって算出された 2 つの語義間の類似度がこの閾値 T_m 以下のとき，それら 2 つの語義は類似していないとみなして対応付けを行わないといった処理をする際に用いられる．ここで T_m は以下のように決定した．テストデータにおける全ての単語の全ての語義の組み合せについて語義間の類似度を計算する．語義間の類似度の分布を図 4.1 に示す．グラフの X 軸は語義間の類似度 (語義ベクトルのコサイン類似度) であり，Y 軸はその出現頻度である．類似度の分布が正規分布にしたがうと仮定し，$[-\infty, T_m]$ の範囲の確率密度関数の確率の累積が全体の 20%になるように (全体の語義の組のうち 20%が対応付け不可となるように) 定めた．図 4.1 にある赤い縦の破線が以上の方法で設定した T_m を表す．実験に使用したテストデータでは，$T_m = 0.196$ と決定した．

同様に，語義間の類似度を WMD で測る場合についても閾値 T_m を設定した．WMD は非類似度を測るため，値が小さければ小さいほど，文間の距離が近くなる．WMD による語義間の距離が閾値以上であれば，2 つの語釈文は類似していないとみなし，対応付けを行わない．テストデータにおける全ての語義間の距離を WMD で測り，その分布を求め，全体の語義の組のうち 20% が対応付け不可になるように閾値 T_m を定めた．その結果，WMD の T_m は 72.613 となった．

[1] https://taku910.github.io/mecab/

表 4.1: テストデータ

難易度クラス	単語
S	教授, 痛感, 人民, 数学, 原稿, 重大, 以来, 享受, 合成, 提案, 森林, 医学, 一旦, 偶然, 座談, 印刷, 去年, 圧迫, 駆逐, 失礼, 性格, 政策, 事件, 水力, 最初, 細胞, 真理, 電信, 診療, 強制, 完全, 協力, 空想, 共同, 採用, 四季, 水産, 勤勉, 以内, 協定, 水運, 作品, 成功, 通商, 財産, 催促, 印象, 教養, 意義, 距離, 巨額, 郷愁, 禁止, 説明, 天才, 伝統, 服装, 遺憾, 通行, 災害, 店員, 教師, 一生, 強硬, 金融, 夫人, 殺人, 以後, 競争, 支持, 一斉, 空中, 死刑, 停滞, 天然, 議論, 試験, 不快, 才能, 服従, 銀行, 緯度, 新郎, 以前, 産地, 苦心, 関心, 転換, 夫妻, 散歩, 定価, 新年, 金属, 訂正, 資金, 作家, 空襲, 停止, 作文, 外部, 安全, 業績, 提供, 認識, 巨大, 人類, 課題, 随意, 睡眠, 転化, 進歩, 電灯, 性質, 愛好, 移動, 近年, 電車, 謹慎, 区域, 異議, 意識, 適用, 定義, 通知, 維新, 色彩, 区分, 負傷, 漁業, 事故, 極端, 正式, 程度, 最近, 政治, 資源
O	進行, 信用, 人間, 往復, 先生, 合計, 保険, 尋常, 簡単, 差別, 大事, 電気, 雑誌, 解放, 一切, 検討, 指揮, 裁判, 是非, 強調, 東西, 料理, 言語, 抵抗, 通用, 深刻, 左右, 一定, 徹底, 圧力, 推移, 伝達, 水面, 単位, 通例, 結構, 大抵, 成就, 不平, 貧乏, 緊張, 参加, 教室, 趣味, 意外, 過去, 事業, 一致, 機関, 開発, 空気, 電報, 意見, 伝道, 産業, 記録, 生活, 一刻, 近代, 清潔, 左翼, 多少, 到底, 無理, 数字, 推進, 人物, 一時, 順序, 具体, 勉強, 便宜, 神秘, 大家, 通信, 集中, 光景, 適当, 作用, 自己, 得意, 生産, 対象, 模様, 容易, 分解, 一面, 掃除, 精神, 時刻, 不幸, 曖昧, 参考, 専門, 一般, 第一, 外人, 自覚, 椅子, 行政, 武士, 天気, 以外, 西欧, 始終, 調子, 資格, 恐怖, 異様, 動員, 翻訳, 以下, 招待, 賛成, 下流, 時間, 材料, 事情, 小児, 所有, 心理
D	勤労, 無論, 野菜, 工夫, 行事, 社員, 気味, 新聞, 迷惑, 漢語, 階段, 大丈夫, 大名, 正月, 記事, 用意, 一身, 評判, 汽車, 一向, 講義, 混雑

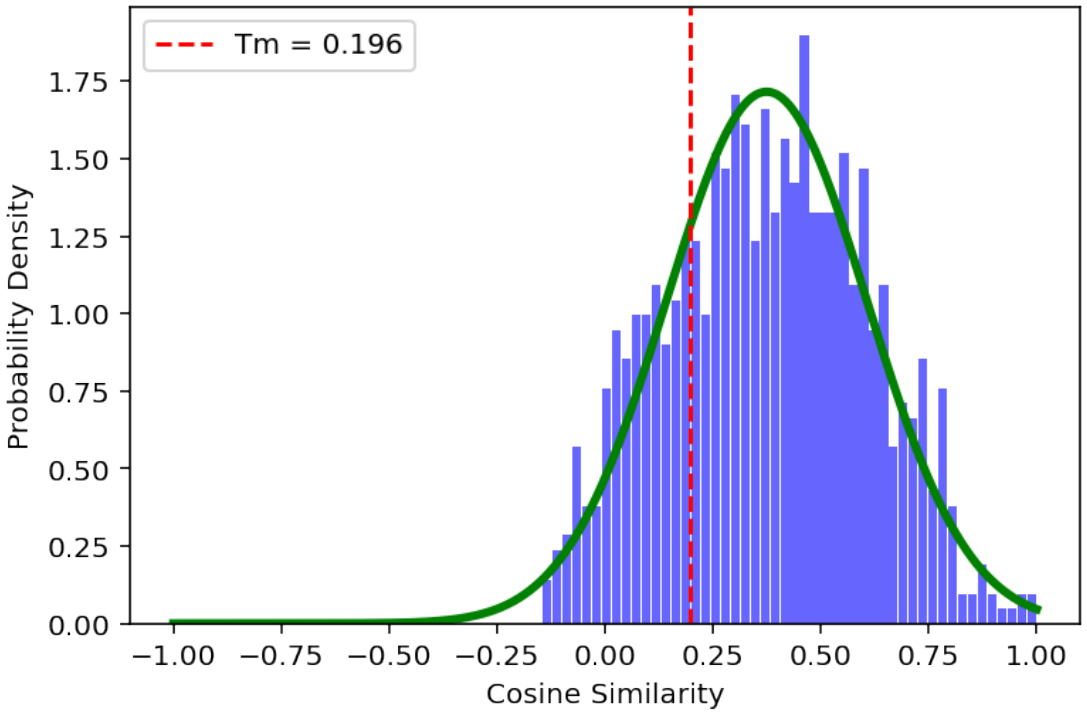


図 4.1: 語義間の類似度の分布

4.1.3 実験結果と考察

各手法の S, O, D の判定の正解率を表 4.2 に示す．ここでは (1) から (5) までの 5 つの判定モデルの結果を載せている.「類似度計算手法」の行は，語義間の類似度を測る手法を指し,「VEC」は語義のベクトル間類似度を用いる手法,「WMD」は Word Mover's Distance を用いる手法を表す.「品詞を考慮」の行は，3.4.3 項で述べた品詞を考慮した語義の対応付けを行うか否かを表す.「多対多を考慮」の行は，3.4.4 項で述べた多対多の対応関係を考慮した語義の対応付けであるか否かを表す.

まず，実験設定 (1) のときの正解率が 0.76 であることから，提案手法による難易度判定はある程度妥当であると言える．さらに，語義の対応付けを行う際に日本語単語と中国語単語の品詞の一致も考慮することによって，正解率は 0.763 まで上昇した．差は小さいが，品詞を考慮することの効果が確認された.

次に，判定モデル (3)，すなわち多対多の対応付けを考慮した難易度判定手法について考察する．この手法では，総スコアが最大となる対応付けの組み合わせを決めるので，閾値 T_m は必要としない．また，この手法は全ての語義の組み合わせに対してアライメントのスコアを計算するため，語義数が多くなると計算量が組み合わせ的に増大し，メモリ不足のために計算が完了しない．そのため，語義数が特に多い 7 つの単語 (「以前」,「一般」,「時間」,「過去」,「多少」,「工夫」,「先生」) を評価用単語から除外した．表 4.2 で判定モデル (3) のテスト単語数が 272

表 4.2: 単語の難易度推定の実験結果

判定モデル	(1)	(2)	(3)	(4)	(5)
類似度計算手法	VEC			WMD	
品詞を考慮	×	○	○	×	○
多対多を考慮	×	×	○	×	×
T_m	0.196		–	72.613	
正解語数	212	213	167	186	188
テスト単語数	279		272	279	
正解率	0.760	**0.763**	0.614	0.667	0.674

になっているのはそのためである．判定モデル (3) の正解率は 0.614 となり，多対多の対応関係を考慮しない判定モデル (1) や (2) の正解率を下回る結果となった．

最後に，語義間の類似度を WMD で測るモデルについて考察する．表 4.2 に示すように，品詞を考慮しない場合 (判定モデル (4)) と考慮した場合 (判定モデル (5)) の正解率はそれぞれ 0.667，0.674 となった．これらの結果は，語義間の類似度を語義ベクトルのコサイン類似度で測る判定モデル (1) や (2) よりも悪かった．この原因として，日本語単語の語釈文と中国語単語の語釈文とで文の長さに大きな差がある場合が多く，このような場合に WMD では文間の距離を正確に測ることができなかったことが考えられる．

5 つの判定モデルの中で正解率が最大となったのは判定モデル (2) であった．判定モデル (2) の対応表を表 4.3 に示す．行の S, O, D は正解の難易度クラスを，列の S, O, D は実験による判定結果を表す．例えば，正解が S クラスの 136 語を提案手法によって判定した結果，106 語を S クラスに，26 語を O クラスに，4 語を D クラスに判定した．「一致数」は S, O, D のそれぞれについて，正解とシステムの判定結果が一致した単語数を示す．「再現率」は，難易度判定の再現率，すなわち正解が S(または O または D) クラスである単語のうちシステムによって難易度を正しく当てることができた単語の割合を示す．S クラスと O クラスに関しては，再現率はそれぞれ 0.779 と 0.851 であり，良好な結果が得られたことが分かった．その一方で，D クラスの再現率はわずか 0.182 で，D クラスについてはほとんど正しく分類できなかった．多対多の対応関係を考慮せず，類似度の高い語義の組から順に語義の対応付けを決定する手法では，D クラスの単語の検出に弱いことが判明した．

次に，多対多の対応付けを考慮した判定モデル (3) の対応表を表 4.4 に示す．判定モデル (2) では D クラスの再現率は 0.182 と低かったが，判定モデル (3) では 0.333 に改善されている．また，S クラスについても，判定モデル (3) の方が再現率が高い．一方，O クラスについては再現率が 0.851 から 0.336 と大きく低下した．判定モデル (3) では，全ての単語に対して S クラスに判定する傾向が強い．このことは，システムが S クラスに分類した単語の合計が，表 4.3 では 133 である

表 4.3: 判定モデル (2) の対応表

難易度	S	O	D	計	一致数	再現率
S	106	26	4	136	106	0.779
O	18	103	0	121	103	0.851
D	9	9	4	22	4	0.182
計	133	138	8	279	213	0.763

表 4.4: 判定モデル (3) の対応表

難易度	S	O	D	計	一致数	再現率
S	121	6	8	135	121	0.896
O	71	39	6	116	39	0.336
D	13	1	7	21	7	0.333
計	205	46	21	272	167	0.614

表 4.5: 判定モデル (3)+(2) の対応表

難易度	S	O	D	計	一致数	再現率
S	103	23	9	135	103	0.763
O	15	95	6	116	95	0.819
D	8	6	7	21	7	0.333
計	126	124	22	272	205	0.754

のに対し，表 4.4 では 205 に増えていることから確認できる．これにより S クラスの再現率は上がったが，O クラスの再現率は低下したと考えられる．

そこで，判定モデル (2) と (3) を組み合わせたモデルを評価した．具体的には，まず判定モデル (3) で難易度を推定し，S クラスと判定したとき，判定モデル (2) で難易度を再判定した．以下，これを「判定モデル (3)+(2)」と呼ぶ．このモデルの対応表を表 4.5 に示す．全体の再現率 (これは正解率に等しい) は 0.754 となり，判定モデル (3) の 0.614 を上回る結果が得られた．判定モデル (2) の正解率 0.763 にはわずかに及ばなかったが，D クラスの再現率は 0.151 ポイント上昇した．

4.1.4 パラメタ T_m 最適化

4.1.2 項で述べたパラメタ T_m は最適ではない可能性がある．テストデータ全体の語義の組のうち 20% が対応付け不可になるように閾値 T_m を決めているが，20% という値は直観で決めている．そこで，3 分割交差検定によって T_m を最適化する実験を行った．まず，テストデータを 3 つのデータに等分割し，3 分の 2 を開発データ，3 分の 1 をテストデータとする．閾値 T_m を 0 から 0.4 の範囲で 0.01 刻みで，0.4 から 1 の範囲で 0.1 刻みで変動させ，開発データにおける正解率が最大と

表 4.6: 実験結果

	(a)	(b)			
	ALL	CV1	CV2	CV3	平均
T_m	0.196	0.34	0.23	0.23	–
正解率	0.763	0.731 (0.769)	0.785 (0.758)	0.753 (0.774)	0.756 (0.767)

なるように T_m を決定する．そして，決定した T_m を用いてテストデータを評価する．テストデータと開発データの入れ換えを 3 回行い，3 つのテストデータの正解率のマイクロ平均を算出する．

実験結果を表 4.6 に示す．(a) は表 4.2 における判定モデル (2) の結果の再掲である．一方，表 4.6(b) は，交差検定の 3 回の試行について，最適化された T_m，テストデータにおける正解率，開発データにおける正解率 (括弧内の数値) を示している．正解率については 3 回の試行のマイクロ平均も示す．開発データとテストデータの正解率にやや差が見られるが，0.02〜0.04 ポイント程度の差に留まっている．また，T_m は 3 回の試行で大きな違いはない．

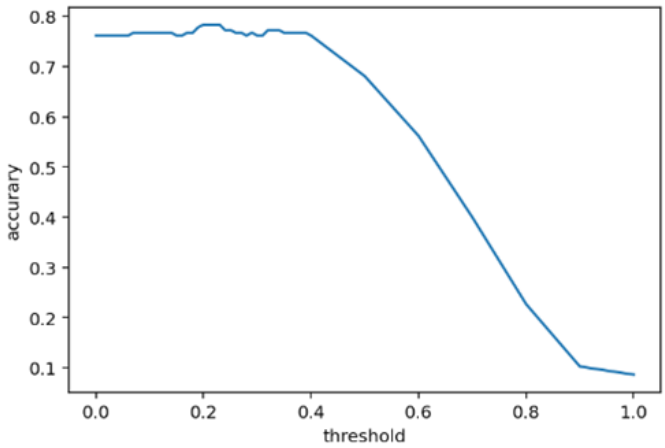
開発データにおいて，閾値を変動させたときの正解率の変化を示すグラフを図 4.2 に示す．このグラフにおいて，X 軸は閾値 T_m，Y 軸は開発データにおける難易度判定の正解率である．3 分割交差検定の 3 回の試行のいずれの場合も，T_m を 0〜0.4 の範囲内で設定したとき，正解率は大きく変わらないことが分かった．以上から，提案手法による判定の正解率は T_m の設定に大きな影響を受けないことが分かる．また，0.196 という値が 0 から 0.4 の範囲内にあることから，4.1.2 項における T_m の決め方は結果として比較的良いヒューリスティクスであったと言える．

4.2 難易度の尺度の妥当性評価

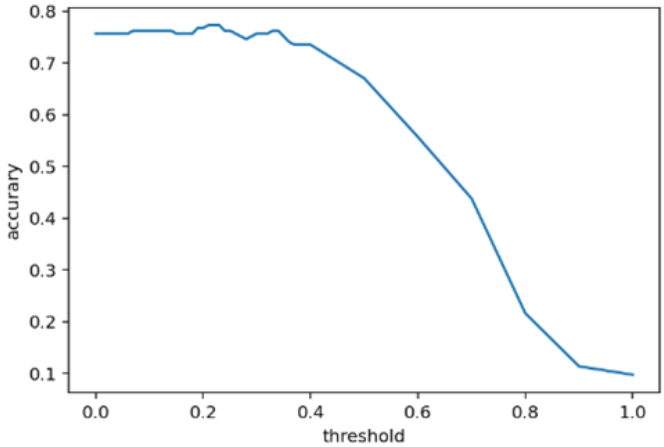
本節は，中国語母語話者を対象としたアンケート調査によって，日本語単語難易度の尺度の妥当性を検証する．被験者に対して日本語単語の難易度を尋ね，この難易度が提案手法の単語難易度とどれだけ一致するかを調べる．4.2.1 項では，調査目的，調査対象，調査内容などアンケート調査の詳細を説明する．4.2.2 項はアンケート調査の結果とそれに対する考察を述べる．

4.2.1 アンケート調査の概要

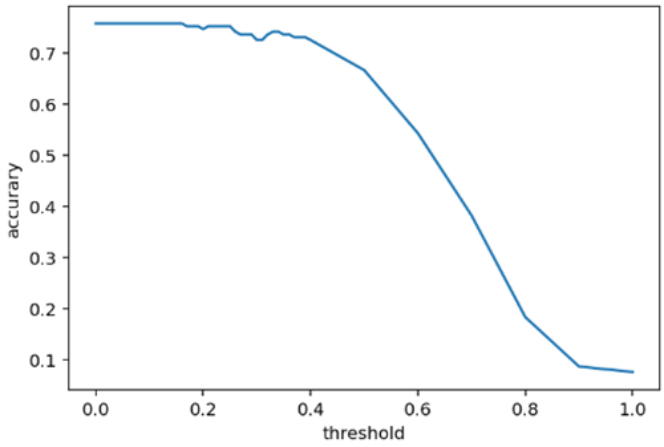
22 名の中国語母語話者を対象に，日本語単語に対して感じた難易度を調べるためのウェブアンケート調査を行った．アンケートで提示した単語は，テストデータから抽出した 20 語を使用した．難易度クラス毎の単語数を表 4.7 に示す．S, O,



(a) CV1



(b) CV2



(c) CV3

図 4.2: 閾値を変動させたときの開発データにおける正解率の変化

表 4.7: アンケートで使用した難易度クラス別の単語数

難易度		単語数	計
S	S-1	2	5
	S-2	3	
O	O-1	2	5
	O-2	3	
D	D-1	2	5
	D-2	3	
N	N	5	5

D, N クラスの単語をそれぞれ 5 語抽出した．ただし，S, O, D クラスに関しては，S-1, O-1, D-1 を 2 語，S-2, O-2, D-2 を 3 語とした．アンケートでは 20 語が 5 つのグループに分けられ，1 つのグループには S, O, D, N クラスの単語がそれぞれ 1 語ずつ含まれるようにした．また，同じグループにおいて S, O, D クラスの単語の漢字表記の難易度は一致させた．つまり，X-1 クラスの単語のあるグループには，X-2 クラスの単語は入れないようにした．例えば，S-1, O-1, D-1, N が一つのグループにある．

アンケートでは，part 1 として，被験者の日本語の学習履歴について以下のように質問する．

1. 日本語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ part 2)

2. どのくらい勉強したか？
   ( ) 年

3. 日本語のレベルは？
   a. ネイティブレベル　b. ビジネスレベル　c. 日常会話レベル　d. 簡単な単語ができる　e. 仮名が読めるだけ

4. 日本語能力試験に合格した場合、そのレベルは？

まず，日本語の学習歴の有無について聞く．これは，初学者とある程度日本語学習の経験がある人 (以下，単に「経験者」と呼ぶ) とで，単語の難易度が変わるかを調べるためである．また，本研究で提案する単語難易度は初めて日本語を学ぶ中国語母語話者を仮定しているが，学習歴のある人は，初学時の単語の難しさを忘れた可能性もある．日本語を勉強したことがない人こそ，初めて日本語を学んだときに感じた難しさを的確に答えられる．次の 3 問は日本語学習の経験の多

さについて質問する．日本語を勉強したことがある人の中には，まだ仮名しか学んでいない人もいる．このような人は，漢字で表記された日本語の単語の勉強をまだ始めていないので，初学者と同じ扱いをするべきである．

次に，part 2 として，日本語の漢字表記語の習得難易度について調査する．各単語に対して，漢字表記，岩波国語辞典の語釈文，例文を提示したあと，単語ごとにその難易度に関する質問をした．被験者には初学者もいるので，語釈文と例文の中国語訳も合わせて提示した．具体的な質問項目を以下に示す．

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない
   語釈 2
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

まず，日本語の学習歴のある人に対しては，その単語を勉強したことがあるかどうかを聞く．単語を勉強したことがある場合は，今覚えている語釈文 (語義) を全て選んでもらう．単語を勉強したことがある人は，習得時の難しさを忘れたかもしれないが，語義を覚えているかどうかは正確に答えられると考えられる．後に詳しく述べるが，覚えていない語義が多いほど，単語の難易度が高いと仮定する．次に，全ての回答者を対象に，各語釈文 (語義) の習得難易度を 3 段階評価で，単語の習得難易度を 5 段階評価で聞いた．後の考察で主に用いるデータは 5 段階の単語の習得難易度であるが，回答者にとって，いきなり単語の習得難易度を聞いてもなかなか答えられないと考え，語義ごとの習得難易度を先に答えることによって，単語の難易度を正しく答えてもらうことを期待した．最後に，参考として，グループごとに 4 つの単語を難しい順で並べてもらった．

ウェブアンケート調査で使用した調査票を付録 B に示す．付録では質問は日本語で書かれているが，実際には中国語で質問が書かれている．

合計 22 名の中国語母語話者に対してアンケート調査を実施した. アンケート回答者の日本語学習歴は 0〜7.5 年で，完全な初学者から上級者まで網羅的に回答が得られた．表 4.8 にアンケートの回答者の日本語レベルを示す．回答者の日本語レベルは，ネイティブレベルの学習者はいなかったものの，日本語を勉強したことがない人からビジネルレベルの学習者まで，ある程度均等に分布していた．

提案手法の難易度は単語を初めて習得するときの難しさを表すため，経験者よりも初学者にとっての難易度に近いと考えられる．このことを検証するために，回答者の日本語レベルを大きく初学者と経験者の 2 つのクラスに分類した.「勉強したことがない」と「仮名が読めるだけ」は初学者とし,「簡単な単語ができる」,「日常会話レベル」,「ビジネスレベル」及び「ネイティブレベル」は経験者とした．仮名が読めるだけの学習者はまだ本格的に日本語単語の勉強を始めていないので，勉強したことがない人と同じく初学者に分類した．

表 4.8: アンケート回答者の日本語レベル

クラス	日本語レベル	人数	計
初学者	勉強したことがない	6	7
	仮名が読めるだけ	1	
経験者	簡単な単語しかできない	7	15
	日常会話レベル	2	
	ビジネスレベル	6	
	ネイティブレベル	0	

4.2.2 結果と考察

アンケートの part 2 の設問 4 で被験者が 5 段階で評価した各単語の難易度をスコアとして採用し，各単語の平均スコアを算出した．回答の選択肢の a, b, c, d, e はスコアとしてそれぞれ 1, 2, 3, 4, 5 で表す．この平均スコアは各単語の実際の習得難易度を示すと考えられる．提案手法の難易度が実際の習得難易度を表しているかを調べるために，提案手法の難易度と調査結果の平均スコアの相関性の強さを検証した．

相関の強さを調べる 2 つのデータが正規分布でない場合，相関係数は実際の値からではなく，データの順位から計算する必要がある．Spearman の ρ と Kendall の τ はそのために設計された順位相関係数であり，2 つの順位付きのデータが与えられたとき，その順位に基づいて 2 つのデータの統計的な関係を評価する．Kendall の τ は τ-a,τ-b および τ-c の 3 種類がある．τ-a と τ-b は正方形の表 (列と行が等しい，すなわち比較する 2 つのデータのスコアの種類の数が同じとき) に使用される．ただし，τ-b は異なるデータが同順位である場合にも適用できる．提案手法の難易度クラスによる単語の難易度の順位付けは同順位を含むが，τ-b はこのような場合にも適用できる．一方，τ-c は長方形の表 (列と行が等しくない) に用いられる．表が正方形の場合，τ-b と τ-c はほぼ同じである．今回の検証では，調査に用いた 20 個の単語に対して，提案手法による難易度のデータとしてはテストデータにおける 4 段階または 7 段階の難易度を使用し，被験者が感じる難易度のデータとして

表 4.9: 4 段階の難易度とアンケート調査の難易度スコアの順位相関係数

		全回答者	初学者	経験者
Spearman's ρ	相関係数	0.533*	0.763**	0.346
	有意確率 (両側)	0.016	0.0001	0.135
Kendall's τ-c	相関係数	0.453*	0.687**	0.28
	有意確率 (両側)	0.022	0.0005	0.156

表 4.10: 7 段階の難易度とアンケート調査の難易度スコアの順位相関係数

		全回答者	初学者	経験者
Spearman's ρ	相関係数	0.534*	0.788**	0.321
	有意確率 (両側)	0.015	0.00004	0.168
Kendall's τ-c	相関係数	0.408*	0.648**	0.233
	有意確率 (両側)	0.021	0.0002	0.185

はアンケート調査による単語難易度の平均スコアを使用する．平均スコアの種類の数は，提案手法による 4 段階または 7 段階の難易度より数が多い．つまり，両者でスコアの種類の数が異なる．そのため，本研究では順位相関係数の Spearman の ρ と Kendall の τ-c を利用した．

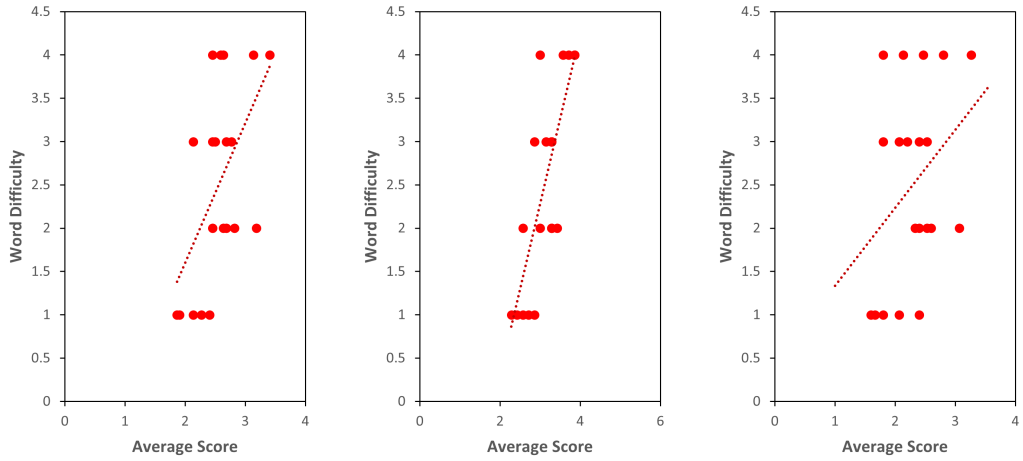
提案手法による難易度を，4 段階の S, O, D, N と 7 段階の S-1, S-2, O-1, O-2, D-1, D-2, N の 2 つのケースに分けて検証する．表 4.9 は 4 段階の難易度クラスについての順位相関係数を，表 4.10 は 7 段階の難易度クラスについての順位相関係数を示している．なお，記号「*」は，順位相関係数が両側検定 5% 水準で有意であることを，記号「**」は両側検定 1% 水準で有意であることを示している．

表 4.9 より，全回答者では，4 段階の難易度とアンケート調査による難易度の平均スコアとの順位相関係数は，Spearman の ρ が 0.533，Kendall の τ-c が 0.453 で，2% の有意水準でやや相関があることが分かった．さらに，初学者と経験者に分けたときの順位相関係数を見ると，初学者では Spearman の ρ が 0.763，Kendall の τ-c が 0.687 で，かなり強い相関があることが判明した．その一方で，経験者については有意ではなかった．

7 段階の難易度は漢字表記の違いも考慮して定義されている．表 4.10 より，7 段階の難易度とアンケート調査による難易度の平均スコアとの順位相関係数は，全回答者では，Spearman の ρ が 0.534，Kendall の τ-c が 0.408 で，4 段階の難易度と同じく 2% の有意水準でやや相関がある．また，初学者では，Spearman については相関係数が 0.788 で有意確率が 0.00004，Kendall については相関係数が 0.648 で有意確率が 0.0002 となり，どちらの検定でも相関が高いことが分かった．一方，経験者では 4 段階の難易度と同じく有意な結果が得られなかった．

経験者に限ったときには順位相関係数が低い，すなわち経験者が感じる単語難

易度と提案手法で定義された単語難易度が一致しない原因を考察する．図 4.3 は，提案手法による 4 段階の難易度と，アンケート調査で得られた単語難易度の平均スコアとの関係をグラフとして可視化したものである．グラフの X 軸は平均スコアであり，Y 軸は提案手法の 4 段階の難易度である．難易度クラスの S, O, D, N はそれぞれ 1, 2, 3, 4 で表す．この図から分かるように，初学者では $y = x$ の直線近くにデータがプロットされているのに対して，経験者に関しては直線から外れた点がたくさん見られる．特に，ほとんどの O クラスの単語については，D クラスの単語より，被験者が回答した難易度の平均スコアが高かった．この結果から，経験者が感じる単語の難易度は日本語単語と中国語単語の語義の類似性ではなく，それとは別の要素に影響されたのではないかと推測した．



(a) 全回答者　(b) 初学者　(c) 経験者

図 4.3: 4 段階の難易度と調査結果の平均スコアとの関係

ここで，経験者が感じる単語の難易度は，以前に学習した単語の意味をどれだけ覚えていられるかに影響するという仮説を立てる．この仮説を検証するために語義忘却率 (forgetting-rate) を用いる．語義忘却率とは，習得した単語の全ての語義の数に対して，覚えていない語義が占める割合と定義する．具体的には，語義忘却率は式 (4.1) のように定義した．単語 w の語義の集合を $S^w = \{s_1^w, ..., s_i^w, ..., s_n^w\}$ とし，n 個の語義があるとする．L^w は単語 w を勉強したことのある人数で，R_i^w は単語 w の i 番目の語義を覚えている人数である．アンケート調査では，調査票の part2 の設問 2 において，その単語を知っている人に語義を覚えているかを尋ねているので，この結果から L^w や R_i^w を求めることができる．

$$\text{forgetting-rate} = 1 - \frac{\sum_{i=1}^{n} R_i^w}{L^w * n} \tag{4.1}$$

S, O, D, N による単語難易度と語義忘却率の関係を図 4.4 に示す．グラフの X 軸は提案手法の 4 段階の難易度であり，Y 軸は各難易度の単語の語義忘却率であ

る．4 段階の難易度クラスの S, O, D, N は X 軸においてそれぞれ 1, 2, 3, 4 で表す．O クラス $(X = 2)$ においては，D クラス $(X = 3)$ の単語より語義忘却率が高い単語が 4 つあることが分かった．すなわち，語義忘却率は D クラスよりも O クラスの単語の方が全般に高かった．一方，先に述べたように，経験者については，O クラスの単語に対する難易度スコアの平均は D クラスよりも大きかった，つまり経験者は O クラスの単語が D クラスの単語よりも難しいと感じていた．この原因として，語義忘却率が大きいと単語を難しく感じる可能性がある．

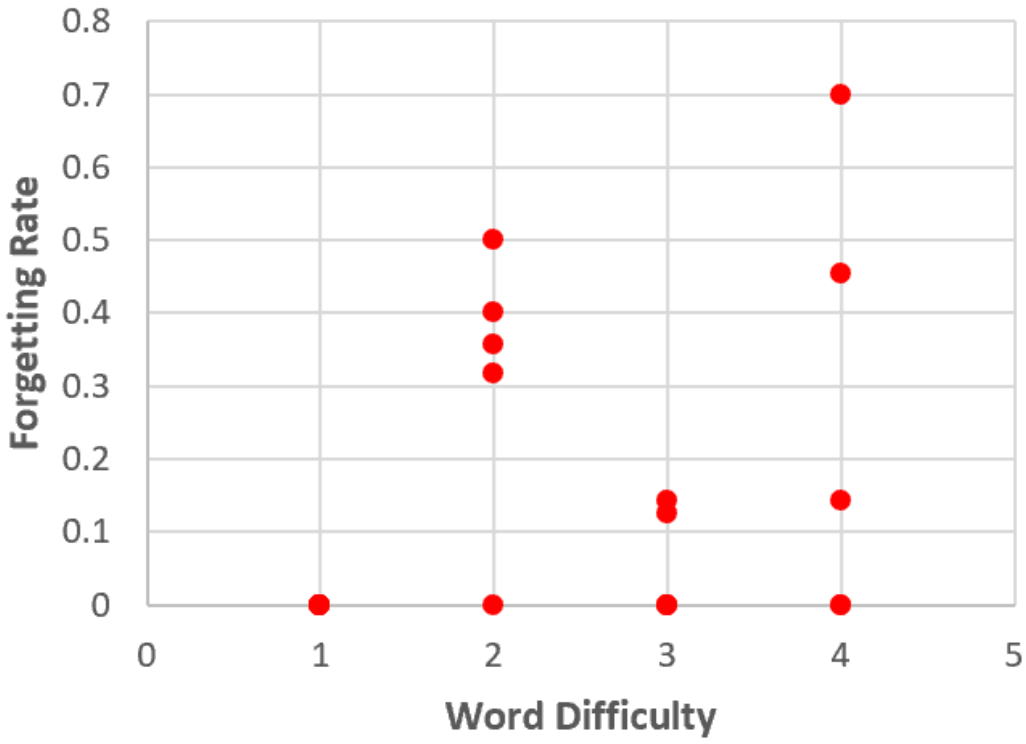


図 4.4: 各難易度の単語の語義忘却率

単語忘却率と単語の難易度の関係をさらに検証するために，単語を勉強したことのある人に対して，難易度の平均スコアと単語の語義忘却率との順位相関係数を計算した．その結果を表 4.11 に示す．単語の語義忘却率と単語難易度の平均スコアは相関が強いことが判明した．つまり，経験者にとっての難易度は語義忘却率と，つまり勉強した後に語義を忘れてしまったかどうかと関係がある．勉強をしてからある程度時間が経ったあと，忘れた語義が多ければ，単語に対して語義が覚えにくいという印象を持つようになり，単語の難易度も高く感じると考えられる．アンケートに用いた O クラスの単語の語義の数は少なくとも 2 つあり，使用頻度が低い語義は忘れやすいので，経験者が感じた難易度もそれに応じて高くなったと考えられる．

厳密な結論を得るには更なる検証が必要だが，経験者が感じる単語の難易度が

表 4.11: 単語の語義忘却率と平均スコアの順位相関係数

		単語既修者
Spearman's ρ	相関係数	0.742**
	有意確率 (両側)	0.00002
Kendall's τ-c	相関係数	0.622**
	有意確率 (両側)	0.0004

提案手法の難易度と一致しないのは，経験者にとっての難易度が，日中の単語の語義の類似性だけでなく，語義忘却率によって影響を受けるからであると言える．提案手法の単語の難易度は，経験者にとっての難易度の指標としては適切ではない．しかし，本研究では主に初学者を想定し，単語を初めて勉強したときの難易度を推定することを目的としている．表 4.9 や表 4.10 の結果から，このことがある程度達成できたことが確認された．

第5章　おわりに

5.1　まとめ

本論文では，漢字を知っている中国語母語話者から見た日本語の単語の難易度を自動的に推定する手法を提案した．

提案手法では，資料「中国語と対応する漢語」における分類を基に，日本語の単語 (漢字で表記された単語に限る) の難易度を，難易度が低い順に，大きく S(日中における意味が全て同じ)，O(日中における意味が一部重なっている)，D(日中における意味が著しく異なる)，N(日本語と同じ単語が中国語に存在しない) の 4 つのクラスと定義した．また，クラス S, O, D については，日本語単語と中国語単語の漢字表記が一致しているときは X-1，異なるときは X-2 のように難易度をさらに細分化した．

S, O, D, N の分類は以下のように行った．まず，日本語の漢字表記語と同じ単語が中国語の辞書にあるかを検索した．もし見つからない場合，日中漢字マッピングテーブルを参照し，日本語の漢字を中国語の漢字に変換してから検索した．漢字を変換しても見つからない場合，N と分類した．それ以外は，日本語辞書から日本語単語の語釈文を，中国語辞書から中国語単語の語釈文を抽出した．日本語の辞書として岩波国語辞典を，中国語の辞書として白水社中国語辞典と現代漢語詞典を用いた．次に，日中の辞書の語釈文の類似度を計算し，似ている語義の対応付けを行った．語釈文のベクトルは，語釈文に含まれる単語の分散表現の平均ベクトルとし，日中の辞書の語釈文の類似度は，2 つの語釈文のベクトルのコサイン類似度で表した．日本語の語義と中国語の語義の全ての組み合わせのうち，類似度が最も高くなる語義の組から対応付けを行った．対応付けられた語義の組を除き，残りの語義の組について同じ処理を繰り返した．語義間の類似度が閾値 T_m 以上ではないとき，二つの語義は同じ意味を持つとはみなさず，語義の対応付けを終了した．語義の対応付けの終了後，対応付けできる語義の組が 1 つも見つからないときは D と判定し，対応付けできる語義の組はあるが全ての語義について対応付けできないときは O と判定し，全ての語義について対応付けができたときは S と判定した．中国語の辞書として白水社中国語辞典と現代漢語詞典を用いたときのそれぞれについて上記の手法を適用し，2 つの辞書による判定結果が異なる場合は，語義の対応付けのスコア (対応付けられた語義間の類似度の平均値) を算出し，大きい方の判定結果を採用した．以上で述べた基本手法に加え，品詞を考慮した手法と多対多を考慮した手法も提案した．

提案手法の有効性について 2 つの方法で評価した．一つ目は，提案手法による S, O, D の難易度クラスの判定の性能を評価した．人手で正解の難易度を付与した 279 語の漢字表記語をテストデータとし，提案手法による難易度判定の正解率を調べた．日中の語義間の類似度の分布が正規分布にしたがうと仮定し，$[-\infty, T_m]$ の範囲の確率密度関数の確率の累積が全体の 20%になるように閾値 T_m を定めた．その結果，T_m を 0.196 と設定したとき，品詞を考慮した手法の正解率が 0.763 となり，全ての実験設定の中で最も高い正解率であった．以上のことから，提案手法による難易度の判定はある程度妥当であると言える．3 分割交差検定によって閾値 T_m を最適化したところ，0.196 に近い値になったことから，実験での T_m の決め方は結果として比較的良いヒューリスティクスであったことが分かった．

二つ目の評価として，中国語母語話者を対象としたアンケート調査によって，日本語単語難易度の尺度の妥当性を検証した．テストデータから S, O, D, N クラスの単語をそれぞれ 5 語抽出した．22 人の中国語母語話者に対して，これらの 20 語の日本語単語の難易度を 5 段階で評価することを依頼した．これらの単語に対する提案手法による難易度とアンケート調査によって得られた難易度の平均値との順位相関係数を調べた．その結果，被験者を日本語学習の初学者に限ったとき，両者には 1% の有意水準で相関があることが判明し，提案手法の尺度の妥当性が確認された．

5.2 今後の課題

本論文の今後の課題を述べる．大きく分けて 3 つの課題があると考えている．

一つ目の課題は語義間の類似度計算の精度を上げることである．本論文では文に含まれる単語のベクトルの平均ベクトルを文ベクトルとし，コサイン類似度によって語義間の類似度を計算した．WMD という手法も取り入れてみたが，今回の実験ではその有効性は確認できなかった．文の分散表現として，Sentence BERT など，ほかの最新の技術を試して類似度を計算し，語義の対応付けの精度を上げる必要がある．また，本論文では中国語で書かれている現代漢語詞典の語釈文との類似度を計算するとき，翻訳 API を用いて翻訳し，同じ言語の文間の意味的類似度を計算した．しかし，正しく翻訳できない文も多々あった．異なる言語の文間の意味的類似度計算の手法を用いることで，翻訳の誤りに影響を受けず，正確に語義間の対応付けを行うことができるかどうかを検討する必要がある．

二つ目の課題は多対多を考慮した語義の対応付けアルゴリズムの改良である．本論文では全ての語義の組み合わせに対して，対応付けられたときのスコアと対応付けないときのスコアを計算し，その和が最大となる場合の組み合わせを求めた．しかしながら，評価実験ではこの手法の正解率は低かった．閾値を用いる手法と組み合わせることで正解率が上昇したが，十分に高い正解率が得られたとは言えなかった．多対多の対応を考慮した語義の対応付けは複雑で難しい問題ではあるが，提案手法をさらに洗練し，より良い手法を探究することが重要な課題となる．

三つ目の課題は難易度の定義の見直しである．まず，O クラスに関して，本論文では一部の意味が重なっていれば全て O クラスと判定したが，重なっている語義の一般性，すなわち語義の日中における使用頻度によって，難易度が大きく変わると考えられる．重なった語義の使用頻度が大きいと難易度が低くなる．したがって，現在の難易度の定義で O クラスに分類される単語でも難易度に大きな差があると考えられる．しかしながら，語義の使用頻度が分かるデータがなかったため，語義の使用頻度は単語の難易度の定義に取り入れなかった．解決策として，例えば，日本語コーパスや中国語コーパスにおける単語の語義を自動的に決定し，語義の出現頻度を見積ることが考えられる．次に，N クラスに関して，本論文では中国語に存在しない単語は N クラスと判定したが，N クラスの単語でも中国語母語話者が単語の意味を推測できるものと推測できないものがある．推測可能性を測る手法として，中国語における漢字ごとの語義を合わせて，日本語単語の語義との類似度を計算することが考えられる．この手法は，漢字表記語ではなく，漢字の当て字のある和語に対しても適用できるのではないかと考える．最後に，漢字表記による難易度の細分化も見直しが必要である．本論文では漢字が完全に一致している場合と異なる場合によって難易度を区別したが，中国語と日本語の漢字が異なる程度にも差がある．例えば，前に例として示した「議論」と「议论」のように漢字の字体が大きく違うものもあれば,「単」と「単」のように字体の差がわずかなものもある．字体の違いの程度を測って単語の難易度を推定することは検討に値する．

参考文献

[1] 浅原正幸, 岡照晃. nwjc2vec:『国語研日本語ウェブコーパス』に基づく単語の分散表現データ. 言語処理学会第 23 回年次大会発表論文集, pp. 94–97, 2017.

[2] Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. Archiving and analysing techniques of the ultra-large-scale web-based corpus project of NINJAL, Japan. *Alexandria: The Journal of National and International Library and Information Issues*, Vol. 25, No. 1-2, pp. 129–148, 2014.

[3] Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. Chinese characters mapping table of Japanese, traditional Chinese and simplified Chinese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pp. 2149–2152, 2012.

[4] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015.

[5] 水谷勇介, 河原大輔, 黒橋禎夫. 日本語単語の難易度推定の試み. 言語処理学会第 25 回年次大会発表論文集, pp. 670–673, 2018.

[6] 中西聖明, 木藤善信, 木村祐介, 椎名広光, 北川文夫. 日本語の単語難易度推定による VOD 講義の難易度推定. 研究報告データベースシステム (DBS), Vol. 2011-DBS-153, No. 8, pp. 1–8, 2011.

[7] Institute of Linguistics, Chinese Academy of Social Sciences. Contemporary Chinese Dictionary (現代漢語詞典), 5th edition. The Commercial Press, 2005.

[8] 劉志宇, 内田理. 日本語を学習する外国人を対象とした日本語テキスト難易度推定手法. 研究報告自然言語処理 (NL), Vol. 2012-NL-205, No. 11, pp. 1–5, 2012.

[9] 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.

[10] 玉岡賀津雄. 中国語と英語を母語とする日本語学習者の漢字および仮名表記語彙の処理方略. 言語文化研究, Vol. 17, No. 1, pp. 65–77, 1997.

[11] 沖森直子. 図解 日本の語彙. 三省堂, 2011.

[12] 文化庁. 中国語と対応する漢語. 日本語教育研究資料, pp. 85–143, 1978.

付 録A　語釈文以外の情報を削除するルールの一覧

白水社中国語辞典，現代漢語詞典，岩波国語辞典における語釈文以外の情報を削除するルールを以下に示す．表3.3では分かりやすさのためルールを簡易に記述したが，ここでは正規表現でルールを記述する．

表 A.1: 白水社中国語辞典における語釈文以外の情報を削除するルール

タイプ	ルール
用法	‘.[^\ n]*?’の形で用いる。 〔.[^\ n]*?の形で用いる； 〔.[^\ n]*?の形で用い， 〔.[^\ n]*?の形で， (.[^)]*?用いる。) (.[^)]*?用い) (.[^)]*?用いて) (.[^)]*?用いるが (.[^)]*?用いて (.[^)]*?用い； (.[^)]*?用い， (.[^)]*?用い (^[0-9０-９①②③④⑤⑥⑦⑧⑨]+).[^,。\ n]*?用いて (^[0-9０-９①②③④⑤⑥⑦⑧⑨]+).[^,。\ n]*?用い，
読み	(.{2}方言では[āáǎàēéěèīíǐìōóǒòūúǔùǖǘǚǜüêɑńňǹg a - z A - Z A-Za-z\s・]+) [0-9]\(\(方言\)\).*?\n ([āáǎàēéěèīíǐìōóǒòūúǔùǖǘǚǜüêɑńňǹg a - z A - Z A-Za-z\\s// -]+) [āáǎàēéěèīíǐìōóǒòūúǔùǖǘǚǜüêɑńňǹg a - z A - Z A-Za-z\\s// -]+
その他	◆.[^\ n]*?\ n (～儿) ≧.[^。]*?。 (～) \ (\ (.[^\ n]*?\ )\ ) ⇒.[^\ n]*?。 ≒.[^\ n]*?。 ≦.[^\ n]*?。 ↔.*?。 ≡.*?。 〔.[^〕]*?〕 (.*?\ s) (.*?’を伴い) \ s{2,} (\ pScript=Han) (\ pScript=Han) ／.*?＝.*?。

表 A.2: 現代漢語詞典における語釈文以外の情報を削除するルール

タイプ	ルール
例	例如.*?。([①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳㉑㉒㉓㉔㉕])
	例如.*?$
	，如。
言い換え	(。)[^。①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳㉑㉒㉓㉔㉕]*?所以叫[^。]*?。
	(。)[^。①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳㉑㉒㉓㉔㉕]*?也叫[^。]*?。
	(。)[^。①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳㉑㉒㉓㉔㉕]*?简称[^。]*?。
読み	^[αĀÁÀĒÉÈāáàēéěèēīíìōóǒòūúǔùǖǘǚǜüa - z A - Z A-Za-z•']+\n
	([αĀÁÀĒÉÈāáàēéěèēīíìōóǒòūúǔùǖǘǚǜüa - z A - Z A-Za-z•']+)
姓	【.*?】.*?[αĀÁÀĒÉÈāáàēéěèēīíìōóǒòūúǔùǖǘǚǜüa - z A - Z A-Za-z•']名姓[^名是#]*?。
	【[^。①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳㉑㉒㉓㉔㉕]*?（[αĀÁÀĒÉÈāáàēéěèēīíìōóǒòūúǔùǖǘǚǜüa - z A - Z A-Za-z•']+）名姓[^名是#]*?。
	[①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳㉑㉒㉓㉔㉕123]名姓[^名是#]*?。
	[①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳㉑㉒㉓㉔㉕123]（[αĀÁÀĒÉÈāáàēéěèēīíìōóǒòūúǔùǖǘǚǜüa - z A - Z A-Za-z•']+）名姓[^名是#]*?。
	[①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳㉑㉒㉓㉔㉕123]姓[^名是#]*?。
その他	(。)[^。①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳㉑㉒㉓㉔㉕]*?原产[^。]*?。
	。.*?另见.*?。
	。\n.*?另见.*?。
	另见.*?。
	参看.*?〖[^。]*?。
	参看[0-9]+页[^。]*?。
	〈古〉.*。
	＜古＞.*。
	〈.*?〉
	（~的）
	（~地）
	（~儿）
	（~儿的）
	（图见.*?）
	（后面.*?带.*?）
	（跟“[^（]*?”相对.*?）
	（旧时[^）]*?用语）
	[.*?]
	……
	（）
	（[^（）]*?《[^（）]*?》[^（）]*?）
	＜.*?＞
	（－／／－）
	‖.*?。

表 A.3: 岩波国語辞典における語釈文以外の情報を削除するルール

タイプ	ルール
語義 ID	^[0-9-\ s]+\ n
品詞	<POS>.[^>]+?</POS> \ n <POS>.[^>]+?</POS>
例	「.[^」]*?<EX>.[^」]*?</EX>.[^」]*?」\ s*（.[^）]*?）\ s*。 「.[^」]*?<EX>.[^」]*?</EX>.[^」]*?」\ s*（.[^）]*?） 「.[^」]*?<EX>.[^」]*?</EX>.[^」]*?」\ s*。 「.[^」]*?<EX>.[^」]*?</EX>.[^」]*?」 『.[^』]*?<EX>.[^』]*?</EX>.[^』]*?』 「.[^」]*?<EX0>.[^」]*?</EX0>.[^」]*?」\ s*（.[^）]*?） 「.[^」]*?<EX0>.[^」]*?</EX0>.[^」]*?」\ s*。 「.[^」]*?<EX0>.[^」]*?</EX0>.[^」]*?」 『.[^』]*?<EX0>.[^』]*?</EX0>.[^』]*?』 ’。\ s*例.[^\ n]*?「.*?\ n 。\ s*例\ s*、.[^\ n]*?。 。\ s*例えば.*?。 （\ s*例えば.*?） （\ s*例\ s*、.*?） 「.[^」]*?」\ s*\ n 。\ s*「.[^」]*?」\ s*。 「.[^」]*?」\ s*（.[^）]*?）\ s*[^」]\ n
読み	<RB>.[^>]*?</RB> 。.[^。\ n]*?「.[^」]*?」\ s*と\ s*読む\ s*。\ s*\ n 。.[^。\ n]*?「.[^」]*?」\ s*と\ s*よむ\ s*。\ s*\ n

タグ	<IT>
	</IT>
	_{
	}
	^{
	}
	<RB>
	</RB>

漢字表記	<KNZ>.*?</KNZ> <GHS>.*?</GHS> <GHK>.*?</GHK>
出典	<REF>.[^>]*?</REF>.*?。
その他	(?=< 派生)[^\ n]*(?=\ n) (?=< 関連 >)[^\ n]*(?=\ n) ⇔.[^。]*?。 ⇔.[^\ n]*?\ n ▽.[^。]*?。\ s*\ n ▽.*?\ n 〔[^〔]*?〕 。\ s*。

付 録B　アンケート調査票

中国語母語話者を被験者とし日本語単語の難易度を尋ねるアンケート調査に用いた調査票を以下に示す.

日本語単語の習得難易度について

Part1 基本情報

1. 日本語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ Part2)

2. どのくらい勉強したか？
   ( ) 年

3. 日本語のレベルは？
   a. ネイティブレベル　b. ビジネスレベル　c. 日常会話レベル　d. 簡単な単語ができる　e. 仮名が読めるだけ

4. 日本語能力試験に合格した場合、そのレベルは？

Part2 日本語の漢字語を習得する難易度について

このパートでは、5 つのグループの日本語漢字語とその語釈を示します。各グループは４つの単語があります。あなたの知識と理解に基づいて、質問に答えてください。

グループ 1

混雑　こんざつ
語釈１：人や物がいっぱい集まって無秩序になること。ごった返すこと。
「〜した通り」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
□ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
語釈 1
a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

布団　ふとん
語釈１：中に綿などを入れ、布地で縫いくるんだ物。すわる時に敷き、また寝具に使う。
「〜した通り」
語釈２：ガマの葉で編み、座禅などに使う円座。

1. (学習歴のある方のみ) この単語を勉強したことがある？
a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
□ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
語釈 1
a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない
語釈 2
a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

汽車　きしゃ
語釈１：(蒸気の力を動力とした) 機関車によって、客車・貨車を引きレールの上を走る車。
「〜に乗る」

1. (学習歴のある方のみ) この単語を勉強したことがある？
a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
□ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

進歩　しんぽ
語釈１：次第によい方、望ましい方へ進み変わって行くこと。
「医学の〜」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

4 つの単語を難しい順で並べてください。

グループ 2

作品　さくひん
語釈１：製作物。主に、芸術活動によって作られたもの。
「シェークスピアの〜」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

利口　りこう
語釈１：頭がよいこと。要領がよいこと。抜け目がないこと。
「〜な人だからそんなことはしない」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

行事　ぎょうじ
語釈１：儀式化して、または一定の計画のもとに、日を決めて行う事柄・催し。
「昨日わが校の創立 80 周年記念の〜が行われた」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

大家　たいか
語釈１：その道に特にすぐれた人。巨匠。重鎮。達人。
「音楽の〜」
語釈２：大きな家。また、富んだ家。家がらの高い家。たいけ。
「〜の出」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない
   語釈 2
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

4 つの単語を難しい順で並べてください。

グループ 3

講義　こうぎ
語釈 1：学問を解説すること。また、その話。特に、大学での (演習・講読・実習以外の) 授業。
「〜に出る」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

尋常　じんじょう
語釈 1：異常なところがなく、ごく普通なこと。
「この問題は〜な方法では解決できない」
語釈 2：すなおなこと。おとなしく、乱れていないさま。殊勝。悪びれず、取り乱さないさま。
「〜に勝負しろ」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない
   語釈 2
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

議論　ぎろん
語釈１：自分の考えを述べたり他人の考えを批評したりして、論じ合うこと。その論の内容。
「その問題は〜に値しないね」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

無駄　むだ
語釈１：役に立たない (余計な) こと。効果・効用がないこと。
「いまさら後悔しても〜だ」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

4 つの単語を難しい順で並べてください。

グループ 4

途方 とほう
語釈１：手段。てだて。
「〜に暮れる」
語釈２：条理。 すじみち。
「あの人は時々〜もないことを言いだす」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない
   語釈 2
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

不幸　ふこう
語釈１：幸福でないこと。ふしあわせ。
「〜に見舞われる」
語釈２：みうちの者に死なれること。
「このたびのご〜なんとも申し上げようがありません」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない
   語釈 2
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

迷惑　めいわく
語釈１：他人のことで、煩わしくいやな目にあうこと。
「他人に〜をかける」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

催促　さいそく
語釈１：早くするように要求すること。
「友人から本を返すよう〜された」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

4 つの単語を難しい順で並べてください。

グループ 5

転換　てんかん
語釈１：物事の方針・傾向などが、今までと別な向きに変わること。また、変えること。
「イメージ～をはかる」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

無論　むろん
語釈１：言うまでもなく。もちろん。
「～彼も来ます」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

箇条　かじょう
語釈１：いくつかに分けてあげた、その一つ一つの条項。
「要求を～にする」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

調子　ちょうし
語釈 1：音楽の節回しや、話し声の、音の高低のぐあい。
「〜が高い」
語釈 2：物事が進んでゆく時の、進行のぐあい、または勢い。
「〜を整える」

1. (学習歴のある方のみ) この単語を勉強したことがある？
   a. ある (→ 2)　　b. ない (→ 3)

2. (単語既習者のみ) 今覚えている語釈文を全て選んでください？
   □ 語釈 1　　□ 語釈 2

3. この単語を勉強するとき、各語釈の習得難易度は？
   語釈 1
   a. すぐ覚えられる　b. 覚えるのに時間がかかる　c. なかなか覚えられない

4. この単語を勉強するのにどれほど難しいと思う？
   a. とても簡単　b. 簡単　c. どちらとも言えない　d. 難しい　e. とても難しい

4 つの単語を難しい順で並べてください。