Vol.2012-IFAT-105 No.11
Vol.2012-NL-205 No.11
2012/1/21

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

劉　志宇[†]　　内田　理[††]

本研究では，語彙の難易度と構文の複雑さからテキストの難易度を推定し，日本語能力試験の受験級の形で提示する手法を提案する．本手法では，読解学習支援システムであるリーディング・チュウ太・語彙チェッカーと構文解析器を用いて，語彙の難易度と構文の複雑さを表す特徴量を日本語テキストから抽出し，難易度推定を行う．難易度の推定式の導出には重回帰分析を用いる．評価実験より，提案手法の有効性が確認できた．

A Method of Difficulty Level Measurement of Japanese Texts for Non-native Learners of Japanese

Zhiyu Liu[†] and Osamu Uchida[††]

This paper proposes a method of difficulty level measurement of Japanese texts by using the difficulty of vocabulary and the complexity of syntax. The estimated difficulty of the text is presented in the form of JLPT's exam level. The JLPT is a standardized criterion-referenced test to evaluate and certify Japanese language proficiency for non-native speakers. In this study, we use the Reading Tutor Reading Learning System and the parser to extract the difficulty of vocabulary and the complexity of syntax form text as the feature quantity. The difficulty is estimated by using those two kinds of feature quantity. The results of the verification experiment indicate the validity of our method.

1. はじめに

近年，日本の諸技術を学ぶ目的で来日する外国人は増えており，日本語を第二言語として必要とする人々が増加傾向にある．現在，コンピューターによる第二言語学習支援の分野において，自然言語処理技術を応用したシステムが研究されている[1]．例えば，読解支援システムや作文添削システム，作文診断システムなどの研究が代表例として挙げられるが，日本語を学習する外国人向けの日本語テキスト難易度判定手法に関する研究はこれまであまり例がない．

日本語学習者が読解を行う際，テキストの難易度は重要である．例えば，言語教育において，学習者の学習段階に応じたテキストから構成される教材を用いることは極めて重要である．また，日本語の試験問題に用いるテキストの難易度も，学習者の習熟レベルに適していることが望ましい．近年，インターネットの普及に伴い，日本語学習者の教材リソースとして，最新のニュース記事などの電子情報を入手することが容易となった．これらの電子情報を難易度によって自動的に分類できれば，日本語の教材や試験問題作成の支援に応用可能である．また，既存の検索エンジンで得られる検索結果の中には様々な難易度の Web ページが混在しているため，理解しやすい Web ページを探すことは困難である場合が多い[2]．テキストの難易度が判定できれば，ユーザの読解力に適したテキストを優先的に提示することが可能となり，検索効率の向上が期待できる．そこで本研究では，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について検討を行う．難易度は日本語能力試験の受験級の形で提示する．

日本語テキストの難易度を決定する要素は語彙，文法，構文などいろいろあり，しかも各々の要素が複雑に絡みあっている．さらに，テキストの内容自体も難易度に大きくかかわってくる．本研究では，難易度を決定する要素として特に重要であると考えられる，語彙の難易度と構文の複雑さを用いてテキストの難易度を推定する．語彙の難易度を表す尺度としては，日本語能力試験の各受験級の語彙の割合を用いる．日本語能力試験の各受験級の語彙の割合を求めるため，読解学習支援システムであるリーディング・チュウ太・語彙チェッカー[3]（以下，語彙チェッカーと省略する）を用いる．また，構文の複雑さを表す尺度としては，二つの文節の係り受け距離を用いる．短い距離を持つ係り受け関係で書かれた文は構文的に優しく理解しやすく，長い距離の係り受け関係が多くある文は難しいと考えられる．このような考え方に基づいて，文章における係り受け距離が，合計 4 カテゴリーに分けた場合の文節の相対頻度を用

[†] 東海大学大学院工学研究科情報理工学専攻
Graduate School of Engineering, Tokai University
[††] 東海大学情報理工学部情報科学科
School of Information Science and Technology, Tokai University

Vol.2012-IFAT-105 No.11
Vol.2012-NL-205 No.11
2012/1/21

いて，構文の複雑さを表す．日本語能力試験の問題集[4-11]をコーパスとして利用し，重回帰分析により，難易度推定式を導出する．評価実験を行ったところ，提案手法の推定精度は 72.2%であった．

2. 関連研究

建石ら[12]は（1）文の平均の長さ（文字数），(2）各文字種（英字，ひらがな，漢字，カタカナ）の連（同一文字種の文字の一続き）の相対頻度，(3）文字種ごとの連の平均の長さ，(4）読点の数の句点の数に対する比，を用いた複数の難易度算定公式を提案している．また，主成分分析により，読みやすさに関係のある成分を見つけ，その計算式を評価式とした．これらの評価式によって求められた値により，複数のテキストの難易度を比較することができる．しかし，その値が具体的にどの程度の難易度に対応するものであるかは明確ではない. 永田ら[13]はリーディングスピート（RS）を用いて文章の読みやすさを評価する手法を提案している．RS とは，単位時間に読むことができる文字数である．しかしながら，RS は個人性が強く，RS をテキスト難易度推定の評価基準として利用することは適切とは言えない．柴崎ら[14] は，小学校の国語教科書（6 学年×3 種類）のコーパスを作成し，(1）1 文の平均文字数，(2）1 文の平均単語数，(3) 1 文の平均係り受け数，(4) 1 文の平均アイディアユニット数，(5）テキスト内の語種の割合，(6）テキスト内の文字種の割合を説明変数とし，学年を従属変数として重回帰分析を行い，テキストの読みやすさを算出する公式を提案している．しかし，国語科以外の教科や他の分野のテキストへの適用については考慮されていない．近藤ら[15]は，円滑な情報伝達を実現することを目的として，日本の小，中，高の全学年, 全教科を含む 13 段階の教科書コーパスを用いた日本語テキストの難易度推定手法を提案している．しかし，日本語を母国語とする人を対象にしているため，日本語を学習する外国人のための難易度判定としては不十分であることが指摘されている[16].

3. 難易度推定

本研究では，テキストの難易度基準として，日本語能力試験の受験級を使用する．日本語能力試験は[17] 財団法人日本国際教育支援協会と独立行政法人国際交流基金が主催の，日本語を母語としない人を対象に日本語能力を認定する検定試験である．日本を含め世界 58 カ国・地域（2009 年）で実施され，日本語を母語としない人を対象とした日本語の試験としては最も受験者の多い試験である．そのため，日本語能力試験の受験級は，テキストの難易度基準として直感的にわかりやすく，実用的な基準であると考えられる．

本研究では，難易度推定用データベースのテキスト収集源として，日本語能力試験の問題集を用いる. 日本語能力試験の問題は, 過去の膨大な受験者のデータをもとに，多くの専門家によって作成されている．したがって，日本語能力試験の問題は，その難易度（4 種類の難易度 ; 1 級，2 級，3 級，4 級）が既知であると考えられる．特に，長文読解問題では，新聞，雑誌，説明文，手紙など様々なテキストが用いられており，受験級に応じて用いられるテキストの難易度は異なる．例えば，1 級では新聞の論説や評論など，論理的に複雑な文章が用いられるが，2 級以下では，新聞や雑誌の記事，解説など，平易な文章が用いられる．したがって，日本語能力試験の問題は，受験級によってテキストの難易度が異なると考えられる．

このような考えに基づき，本研究では難易度推定用データベースとして，日本語能力試験の問題集[4-11]を収録対象としたデータベースを作成した．具体的には，[4-11]の長文読解問題から設問部分を除いたテキストを用いて電子化データを作成した．難易度推定用データベースの概要を表 1 に示す.

表 1　難易度推定用データベースの概要

	問題集数	サンプル数
1 級	4	36
2 級	4	32
3 級	4	24
4 級	4	20
合計	16	112

4. テキスト難易度と関連する要素

日本語テキストの難易度を決定する要素は語彙，文法，構文などいろいろある．本研究では，テキスト難易度に深く繋がっている語彙の難易度と構文の複雑さを考慮する．

4.1 語彙の難易度とテキスト難易度

難しく馴染みにくい語彙がテキストに出現すると，テキスト難易度は高くなる．例えば，

例文 A（2 級）「スノーボードは魅力的なスポーツであり，広大な大自然を相手に夏山にはない楽しみを私たちに与えてくれる．」

例文 B（3 級）「パソコン室にはパソコンがあるのでここで飲み物を飲まないでください．」

Vol.2012-IFAT-105 No.11
Vol.2012-NL-205 No.11
2012/1/21

というという二つの文に対して，語彙チェッカーを用いて，各日本語能力試験受験級の語彙の数を数えると，2 級の例文 A は一級語 0，二級語 4，三級語 2，四級語 15 となり，3 級の例文 B は一級語 0，二級語 0，三級語 3，四級語 12 となる．このように，受験級によってテキストを構成する各受験級の語彙の割合は異なる．そこで，語彙の難易度を表す尺度としては，日本語能力試験の各受験級の語彙の割合を用いることとする．

4.2 構文の複雑さとテキスト難易度

文を理解するためには，主語と述語との関係や，修飾語と被修飾語の関係などを把握しなければならない．文の構造が複雑になると，例えば，並列構造や長い修飾句などが文中に出現すると，読み返しが行われる．そのため，文の構造が複雑になると，難易度も上昇すると考えられる．日本語においては，主語，述語などの成分で文の構造を表すことが一般的である．本研究では，文の構造を数値化するため，文のすべての成分を文節とそれぞれの係り受け関係で表すことにする．例えば，

例文 C（3 級）　「緑山には美しい湖があって，たくさんの人が遊びに来る．」

という文に対して，構文解析器 CaboCha[18]を適用すると，係り受け関係は図 1 のように表すことができる．

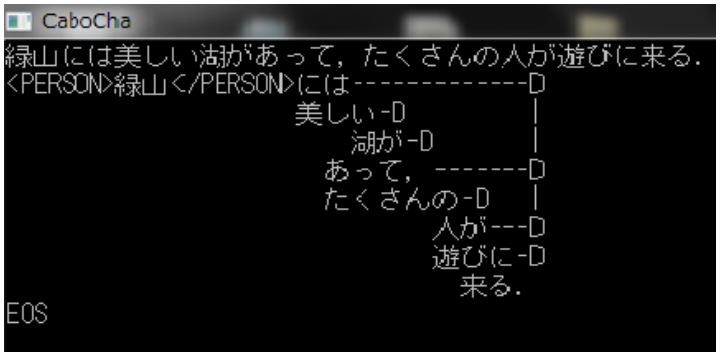

図 1 構文解析器 CaboCha の出力

また，図 1 のような文の構造に基づき，ある文節とその係り先の間に存在する文節の個数に 1 を加えた値を係り受け距離と定義する．上記の例文の文節間の関係を距離別に分けて表すと以下のようになる．

距離 1:
美しい ——— D ——— 湖が
湖が ——— D ——— あって
たくさんの ——— D ——— 人が
遊びに ——— D ——— 来る
距離 2:
人が ——— D ——— 来る
距離 4:
あって ——— D ——— 来る
距離 7:
緑山には ——— D ——— 来る

このように，短い係り受け距離の修飾関係は易しく理解しやすいが，長い係り受け距離をもつ修飾関係は難しい．すなわち，文節間の係り受け距離は構文の複雑さを反映し，テキスト難易度と関わる要素であるといえる．

5. 語彙の難易度と構文の複雑さを表す特徴量の抽出法

本章では，語彙の難易度と構文の複雑さを特徴量としてテキストから抽出する方法を説明する．

5.1 語彙の難易度を表す特徴量の抽出

4.1 で述べたように，語彙チェッカー[3]を用いて，日本語能力試験各受験級の語彙の割合を求める．このシステムには，日日辞書ツールや日英辞書ツール，語彙チェッカー，漢字チェッカーなどを含むが，本研究では語彙チェッカーを用いる．語彙チェッカーは入力された文章に形態素解析を行い，分析結果を日本語能力試験出題基準と照合する．そして，文中の語彙のレベルを日本語能力試験の受験級の形で表示する（図 2）．

総数	語彙総数	級外	1級	2級	3級	4級	その他
610	548	50	26	120	87	265	62
111.3%	100.0%	9.1%	4.7%	21.9%	15.9%	48.4%	11.3%

図 2　語彙チェッカーの出力結果

Vol.2012-IFAT-105 No.11
Vol.2012-NL-205 No.11
2012/1/21

語彙チェッカーの出力結果，すなわち日本語能力試験各受験級の語彙の割合を特徴量として抽出する．例えば，図 2 の場合，1 級語の割合は 4.7%，2 級語の割合は 21.9%，3 級語の割合は 15.9%，4 級語の割合は 48.4%である．

5.2 構文の複雑さを表す特徴量の抽出

4.2 で述べたように，短い係り受け距離の修飾関係で文を書くと文が易しく理解しやすいが，長い係り受け距離を持つ修飾関係を多く使用すると文が難しくなる．構文の複雑さを表す特徴量の抽出法としては，まず，入力された文章を CaboCha を用いて構文解析を行い（図 1），すべての係り受けの距離と係り受けの総数を求める．本研究では，係り受け距離別に 4 つのカテゴリーに分ける．分け方は以下の通りである．

係り受け距離:1~3	カテゴリー1 I
係り受け距離:4~5	カテゴリー2
係り受け距離:6~10	カテゴリー3
係り受け距離:11 以上	カテゴリー4

この 4 つのカテゴリーの係り受けの数と係り受けの総数を求め，それぞれのカテゴリーの係り受けの数を係り受けの総数で割ることにより，カテゴリー別の係り受けの相対頻度を求める．

$$F_i = \frac{n_i}{N}$$

ここで，F_i はカテゴリー i の係り受けの相対頻度，n_i はカテゴリー i の係り受けの数，は N 係り受けの総数である．

この 4 つのカテゴリーの係り受けの相対頻度を構文の複雑さを表す特徴量とする．

6. 重回帰分析を用いた難易度推定式の導出

本研究では，重回帰分析を用いて難易度推定式を導出する．**5.1** と **5.2** で述べた語彙の難易度を表す特徴量，及び構文の複雑さを表す特徴量を説明変数とし，テキストの難易度を目的変数とする．難易度を D で表すとすると，難易度と特徴量の関係は

$$D = \sum_{i=1}^{M} a_i R_i + \sum_{j=1}^{N} b_j F_j + c$$

で表される．ここで，M と N はそれぞれ語彙の特徴量，及び構文の特徴量の種類数である．本研究ではどちらも 4 である．R_i は i 級の語彙の割合である．F_j は j カテゴリーの係り受けの相対頻度を表す．a_i,b_j は回帰係数，c は定数項である．

7. 実験

提案手法の評価実験を以下のように実施した．まず **3.** で述べた難易度推定用データベースから（117 サンプル），各難易度 5 サンプル計 20 サンプルをランダムに選出し，学習用サンプルとした．残り 97 サンプルをテスト用サンプルとした．

次に，**5.** で説明した手法を用いて，学習用サンプルとテスト用サンプルから，それぞれ特徴量を抽出した．20 の学習用サンプルから抽出された特徴量は学習データであり，97 のテスト用サンプルから抽出された特徴量はテストデータである．学習データを用いて，重回帰分析を行い，回帰式を求めた．

$$D = 11.39514R_1 + 7.353623R_2 + 6.604773R_3 + 11.97607R_4$$
$$-2.77517F_1 - 4.06571F_2 - 16.74120F_3$$

ここで，R_i は i 級の語彙の割合，F_j はカテゴリ jーの係り受けの相対頻度を表す．

最後に，以上の二つの回帰式を用いて，テストデータの難易度判定を行った．回帰式の出力を四捨五入により整数に変換し，その整数がテストデータの難易度と等しいかどうかを判定した（ただし，出力が 0.5 未満の場合は 1 級と判定した）．実験結果を表 2 に示す．

表 2 正解率

1 級	2 級	3 級	4 級	全体
80.6%	46.7%	89.5%	82.4%	72.2%

表 2 より，提案手法の有用性が確認できた．

8. おわりに

本研究では，語彙の難易度と構文の複雑さからテキストの難易度を推定し，日本語能力試験の受験級の形で提示する手法を提案した．語彙チェッカーと構文解析器 CaboCha を用いて，語彙の難易度と構文の複雑さを表す特徴量を抽出し，重回帰分析を用いて難易度推定式を導出した．評価実験の結果，難易度の推定精度は 72.2%であり，提案手法が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として有効であることが確認された．

Vol.2012-IFAT-105 No.11
Vol.2012-NL-205 No.11
2012/1/21

本手法では語彙の難易度と構文の複雑さを考慮して難易度推定を行った．しかしながら，文法もテキストの難易度に影響を与える．例えば

例文 D（1 級） 「今週は忙しくて無理だが，来週ならその会に参加できないものでもない．」

という文章に対して，本手法を用いて難易度推定を行ったところ，2 級と推定してしまった．誤推定の理由としては，1 級の文法項目である「～ないものでもない」[19]を考慮していないからと考えられる．今後，文法も考慮することで，更なる難易度推定の精度向上を目指したい．

参考文献

1) V. M.Holland, J. D. Kaplan, and M. R. Sams: Intelligent Language Tutors: Theory Shaping Technology, LEA, pp.183-200 (1995)
2) 中谷誠，アダムヤトフト，大島裕明，田中克己: 理解容易度に基づく Web ページの検索とランキング，電子情報通信学会, (DEIM Forum 2009), A7-1, (2009)
3) 語彙チェッカー: http://basil.is.konan-u.ac.jp/chuta/
4) 凡人社(編): 平成 14 年度日本語能力試験 1・2 級試験問題と正解，凡人社(2003)
5) 凡人社(編): 平成 14 年度日本語能力試験 3・4 級試験問題と正解，凡人社(2003)
6) 凡人社(編): 平成 15 年度日本語能力試験 1・2 級試験問題と正解，凡人社(2004)
7) 凡人社(編): 平成 15 年度日本語能力試験 3・4 級試験問題と正解，凡人社(2004)
8) 凡人社(編): 平成 16 年度日本語能力試験 1・2 級試験問題と正解，凡人社(2005)
9) 凡人社(編): 平成 16 年度日本語能力試験 3・4 級試験問題と正解，凡人社(2005)
10) 凡人社(編): 平成 17 年度日本語能力試験 1・2 級試験問題と正解，凡人社(2006)
11) 凡人社(編): 平成 17 年度日本語能力試験 3・4 級試験問題と正解，凡人社(2006)
12) 建石由佳，小野芳彦，山田ひさお: 日本文の読みやすさの評価式，情報処理学会研究報告, 1988-HI-018, pp.1-8 (1988)
13) 永田亮，井口達也，桝井文人，河合敦夫: リーディングスピードに基づいた文章の読み易さについて，電子情報通信学会技術研究報, TL, Vol.102, No.491, pp.13-18 (2002)
14) 柴崎秀子，沢井康孝: 国語教科書コーパスを応用した日本語リーダビリティー構築のための基礎研究，電子情報通信学会技術研究報告, NLC, Vol.107, No.246, pp.19-24 (2007)
15) 近藤陽介，松吉俊，佐藤理史: 教科書コーパスを用いた日本語テキストの難易度推定，言語処理学会第 14 回年次大会発表論文集, pp.1113-1116 (2008)
16) フメリャク寒川クリスティーナ: 日本語学習者のための日本語テキスト難易度推定用コーパス，電子情報通信学会技術研究報告, TL, Vol.109, No.84, pp.19-24 (2009)
17) 日本語能力試験ホームページ: http://www.jlpt.jp/
18) 日本語構文解析器 CaboCha: http://code.google.com/p/cabocha/
19) 植木香，植田幸子，野口和美: 改定版完全マスター1 級日本語能力試験文法問題対策，スリーエーネットワーク(2006)