

日本語学習者のための文章の難易度判定システムの構築と 運用実験

川村 よし子
東京国際大学

北村 達也
甲南大学

要 旨

本研究の目的は、語彙の難易度と構文の複雑さに着目して、日本語学習者のための文章の難易度レベルを判定するシステムを開発することにある。

川村（2011）において、日本語学習者を対象にした難易度判定実験の結果、文の難易度に最も影響を及ぼすのは語彙の難易度だが、名詞修飾節や重文・複文のような複雑な構文も、難易度を高めることが明らかになった。ところが、文章の難易度の決定に、それぞれがどのような形で影響を及ぼしているかに関する研究は、まだ十分とは言えない。そこで、本研究では、文中の語彙の難易度と一文の長さの平均値を指標に重回帰分析を行い、文章全体の難易度を判定するシステムを構築した。

難易度推定用データベースには、初級・初中級・中級・中上級・上級の各レベルの教科書計 25 冊に加えて、超級（上級以上）のデータとして専門書を利用した。あらかじめ評価実験用データとして抽出したものを除く全データを学習データとして、単語と文の長さを指標に重回帰分析を行い、難易度判定式を求めた。得られた判定式をもとに評価実験を行ったところ、文章の難易度判定に有効であることが確認された。本研究によって開発したシステムは、<http://basil.is.konan-u.ac.jp/chuta/level/>で公開している。

1. 研究の背景

本研究の目的は、語彙の難易度と構文の複雑さに着目して、日本語の文章の難易度レベルを判定するシステムを開発することにある。文の難易度を判定するシステムとしては、川村（1998）が開発し、1998年から Web 上で公開している「語彙チェッカー」がある。これは、旧日本語能力試験

の出題基準（以下「出題基準」）に準拠して、入力文中のすべての単語のレベルを判定するものである。さらに文中の 3 級及び 4 級の単語の比率によって文全体の難易度も表示する仕組みも整えた（川村・北村 1998）。

その後、文の難易度判定システムとして、文中の漢字の難易度と漢字の比率によって文の難易度判定を行う「帯」（Sato, et al. 2008）や、漢字と仮名の比率とともに、文節数と述語の数によって文章の難易度を判定する「日本語リーダビリティ測定」（柴崎 2009）なども公開された。しかし、いずれも文の読み手として日本語母語話者を想定したものであり、日本語学習者のためのものではない。一方、劉・内田（2012）では、語彙チェッカーによる語彙の難易度判定結果と 2 つの文節の係り受け距離を用いて日本語学習者のための難易度推定手法の提案を行っている。評価実験の結果、ある程度の有用性は示されているものの、2 級レベルの教材の判定精度は 50%以下であり、日本語教育にそのまま活用できる状態ではない。その原因として、現在の構文解析の精度の問題がある。そこで、本研究では、文章に含まれる文の複雑さの指標として一文の長さに着目し、文に含まれる単語の難易度と一文の長さの双方を指標として重回帰分析を行い、文章全体の難易度を判定するシステムを構築することにした。

2. 先行研究

2.1 日本語学習者を対象にした難易度判定実験

日本語学習者のための難易度判定システムの構築に先立ち、川村（2011）では、世界の日本語学習者を対象にした難易度判定実験を行った。実験では、語彙の難易度や構文が異なる文を用意し、次の 2 種類の判定を行った。

- a. 文の難易度を 5 段階で評価する（印象による判定）
- b. 文を自身の母語あるいは得意な言語に翻訳する（理解度による判定）

判定結果を分析した結果、文の難易度の決定に最も影響を及ぼすのは語彙の難易度であることが明らかになった。また、特に初級の学習者にとっては、名詞修飾節、テ形および連用中止法の使用等、構文の複雑さや、主格・目的格等の省略が難易度を高める傾向があることが判明した。なお、この実験では語彙の難易度と構文のみに着目し、文の長さに関する調査は行わなかった。

2.2 難易度を高める各要素の検出ツールの開発

学習者を対象にした難易度判定実験の結果を受け、難易度を高める各要素を個別に検出できるツールの開発を行った。水嶋ほか（2011）では、テ形および連用中止法、さらに名詞修飾節の検出ツールを開発した。また、内田ほか（2011）では主格省略を検出するツールを作成し、さらに中村ほか（2012）では、このツールの精度を上げるために検索エンジンを用いた検出方法を提案した。しかし、各ツールの運用実験の結果、いずれのツールの検出精度も、構文解析に用いている構文解析システム KNP（黒橋 2000）の解析精度に影響され、学習支援として活用するには十分とは言えないことが明らかになった。そのため、現時点では構文解析技術を用いることを避け、形態素解析のみを利用した形で難易度判定システムを開発する必要があることが判明した。

2.3 単語と文の長さを基準にした難易度判定システムの開発

一般に、構文が複雑になれば文も長くなり難易度も増す。そこで、川村（2012）では、文に含まれる単語の難易度と文の長さを基準にした難易度判定システムの開発を行った。複雑な文ほど文は長くなる傾向があり、学習者にとっては難しくなる。また、構文が単純であっても長い文は読みやすいとは言えない。そのため、一文の長さの平均値が文章の難易度の指標となりうると考え、語彙の難易度と一文の長さを指標とする難易度判定

ツール B 版を開発した。これは従来の語彙チェッカーの機能に、一文ごとの文字数をカウントし、各文の長さや平均の長さの文字数を示すとともに、色分け表示するという機能を付け加えたものである。文の長さの基準としては、日本語母語話者向けの『介護福祉士新カリキュラム学習ワークブック』の原文と、日本語学習者向けにやさしく書き換えられた文章（川村ほか 2011）における一文の平均文字数を参考にした。日本語母語話者向けの教材の原文では、平均 63.4 字の長さだったのに対し、やさしく書き換えられた文章の一文ごとの長さは、平均 33.3 字だった。そこで、40 字未満の文字数は黒、40～59 字は青、60 字以上は赤で表示することにした。また、入力された文章の一文の平均文字数も同様の基準で表示する仕組みにした。

こうして文の難易度判定ツール B 版が出来上がり、公開を開始したが、色分け表示している長さの基準の妥当性に関しては、多様な種類の文章によって検証していく必要がある。また、文章の難易度に語彙と文の長さのそれぞれがどのようにかかわっているのかを明らかにしなければならない。そのため、今回新たに各レベルの教科書を難易度判定用の学習データとして用いて難易度判定式を求め、それをもとにした難易度判定システムを構築することにした。

3. 重回帰分析を用いた文章の難易度判定

難易度判定システムの開発に先立ち、文章の難易度に対して、単語の難易度と構文の複雑さが各々どのような関係にあるのかを調査する必要がある。単語の難易度の決定に関しては、「語彙チェッカー」同様、出題基準をもとにして行うことにした。また、構文の複雑さに関しては、上述したとおり、構文解析技術にはまだ限界があるため、一文の平均文字数と平均単語数のみを指標に難易度判定を行うことにした。

3.1 難易度判定用学習データベース

難易度判定用学習データベースとしては、初級・中級・上級の各レベルの教科書および専門書と白書（計 28 種）を用意した。いっぽう、開発予定の難易度判定システムは、新しい日本語能力試験や語学のコミュニケーション能力別のレベルを示す国際標準規格である CEFR（Common European Framework of Reference for Languages）、国際交流基金の「JF スタンド」等の学習指標にも対応可能なものを目指している。そのため、日本語能力試験の N1 から N5 にあたる 5 段階に、超級（N0）を加えた 6 段階の判定が行えるように配慮する必要がある。そこで次のような基準で、学習データベースの教科書および専門書・白書を 6 段階に分類した。

N5（A1）：初級教科書の前半

N4（A2）：初級教科書の後半

N3（B1）：中級教科書の前半

N2（B2）：中級教科書の後半（中上級）

N1（C1）：上級教科書

N0（C2）：専門書・白書

学習データベースの本文中に含まれる単語の切り出しには、形態素解析システム MeCab（工藤 2006）を利用した。また、評価実験用データとしては、難易度判定用学習データベースの各々から、あらかじめランダムに 1 章ずつを抽出し、学習データベースとは別に扱うことにした。

3.2 重回帰分析による難易度判定式の決定

上記の教科書および専門書・白書から評価実験用データを除いた全データを、難易度判定用学習データとした。まず、予備調査として、難易度判定のための指標として次の 6 種類を用いて分析を行った。

a. 一文あたりの平均文字数

- b. 一文あたりの平均単語数
- c. 出題基準 1 級の単語数／総単語数（記号を除く）
- d. 出題基準 2 級の単語数／総単語数（記号を除く）
- e. 出題基準 3 級の単語数／総単語数（記号を除く）
- f. 出題基準 4 級の単語数／総単語数（記号を除く）

予備調査の結果、初級・初中級の教材の中に、固有名詞（特に人名）の使用頻度が極端に高いものが含まれているため、難易度判定の分析は固有名詞を除いた形で行う必要があることが明らかになった。また、一文の長さの指標として、[一文あたりの平均文字数]と[一文あたりの平均単語数]のどちらがより有効かを調べるため、各級の単語数の割合に加えて、[平均文字数]を用いた場合と、[平均単語数]を用いた場合とで、学習データをもとに重回帰分析を行った。その結果、前者における[平均文字数]の P-値は 0.002046 であるのに対し、後者における[平均単語数]の P-値は 0.000845 であり、難易度判定の指標としては、[平均単語数]のほうが有効であることが判明した。これは、初級教材では、通常漢字で書かれるべき単語もひらがなで書かれていることが影響していると考えられる。以上の結果から、難易度判定のための指標としては、次の 5 種類を用いることにした。

- a. 一文あたりの平均単語数
- b. 出題基準 1 級の単語数／総単語数（記号および固有名詞を除く）
- c. 出題基準 2 級の単語数／総単語数（記号および固有名詞を除く）
- d. 出題基準 3 級の単語数／総単語数（記号および固有名詞を除く）
- e. 出題基準 4 級の単語数／総単語数（記号および固有名詞を除く）

難易度判定学習データを MeCab を用いて分析し、得られた a から e の各値を表 1 のような形にまとめ、これをもとに重回帰分析を行った。

表 1：難易度判定学習データの分析結果（一部）

学習データ（教科書・白書等）	A	B	C	D	E	F	G
レベル	N0	N0	N0	N0	N0	N1	N1
数値化したレベル	0	0	0	0	0	1	1
a. 一文あたりの平均単語数	26.3	27.8	24.9	29.7	27.1	24.4	20.5
b. 総 N1 数／総単語数（除記号・固有名詞）	6.7	10.7	8.6	9.3	12.6	4.4	4.1
c. 総 N2 数／総単語数（除記号・固有名詞）	19	23.3	23	24.4	20.8	14.1	15.4
d. 総 N3 数／総単語数（除記号・固有名詞）	11.9	10.1	12.8	7.3	8.5	11.7	11.6
e. 総 N4 数／総単語数（除記号・固有名詞）	53.4	47.1	43.4	46.2	43.6	62.1	62.6

重回帰分析の結果、重相関 $R=0.982828$ 、重決定 $R^2=0.96595$ といずれも高い数値で、次のような難易度判定式が得られた。

$$\begin{aligned} \text{レベル } N = & -3.020350224 - 0.108713131 \times [a] + 5.9903188 \times [b] \\ & + 5.3699195 \times [c] + 1.0666679 \times [d] + 9.7980957 \times [e] \end{aligned}$$

判定式の[a]から[e]には、入力された文章における a から e に相当する指標の値が入る。なお、今回の分析では、この 5 つの指標のうち、 $p < 0.01$ となったのは、a [一文あたりの平均単語数]と e [出題基準 4 級の単語数/総単語数（記号および固有名詞を除く）]のみであった。これは、文章の難易度の決定において、一文あたりの平均単語数と 4 級の語彙の比率が他の指標に比べて有意な影響があることを示している。しかし、日本語学習者を対象とした教材作成用の難易度判定システムという性格上、いずれの級の語彙の比率も無視できないものと考え、あえて、a から e のすべてを指標として残した難易度判定式をもとに、評価実験を行うことにした。

4. 難易度判定式の評価実験

3.で得られた難易度判定式が教材の難易度判定に有効かどうかを判定するための評価実験は、難易度判定用学習データとは別の評価実験用データ

をもとに行った。

評価実験用データとしては、上記、難易度判定用学習データとは別に、N5（初級）からN0（超級）までの6レベルの教材から無作為に取り出した教材（各レベル4種類、計24種類）を用いた。実験は、難易度判定システムが評価実験用データの含まれていた元の教科書レベルを正しく評価できるかを調べるという形で行った。評価実験の結果は、次のとおりであった。

- a. 教材レベルと同一判定だったもの：19件
- b. 上のレベルに判定したもの：2件（いずれも1レベルのみのずれ）
- c. 下のレベルに判定したもの：3件（いずれも1レベルのみのずれ）

教材レベルと同一判定だったものは19件（79%）であり、判定がずれた5件（21%）についても1レベルのみのずれであった。また、レベルが上に判定された教材は各教科書の後半部分に置かれた教材であり、レベルが下に判定された教材は各教科書の前半部分の教材であった。こうした結果から、重回帰分析の結果得られた難易度判定式は、かなりの精度でほぼ適切に教材レベルを判定できることが明らかになった。

ただ、今回の実証実験は、あくまでも各教科書が想定している難易度レベルが正しいと仮定しての実験である。このシステムによって判定された難易度レベルが、学習者にとっての文の難易度を正しく示しているかに関しては疑問が残る。難易度判定システムの完成後、システムの示す文章の難易度が、学習者から見た文の難易度と一致するかに関しては、詳しい検証作業を行っていく必要がある。

5. 難易度判定システムの開発

難易度判定式の評価実験の結果、学習データベースをもとに得られた難易度判定式の有効性が明らかになったため、この難易度判定式をもとに文章の難易度判定システムを開発することにした。新しい難易度判定システ

ムは、川村（2012）の難易度判定システムβ版に組み入れる形で開発した。

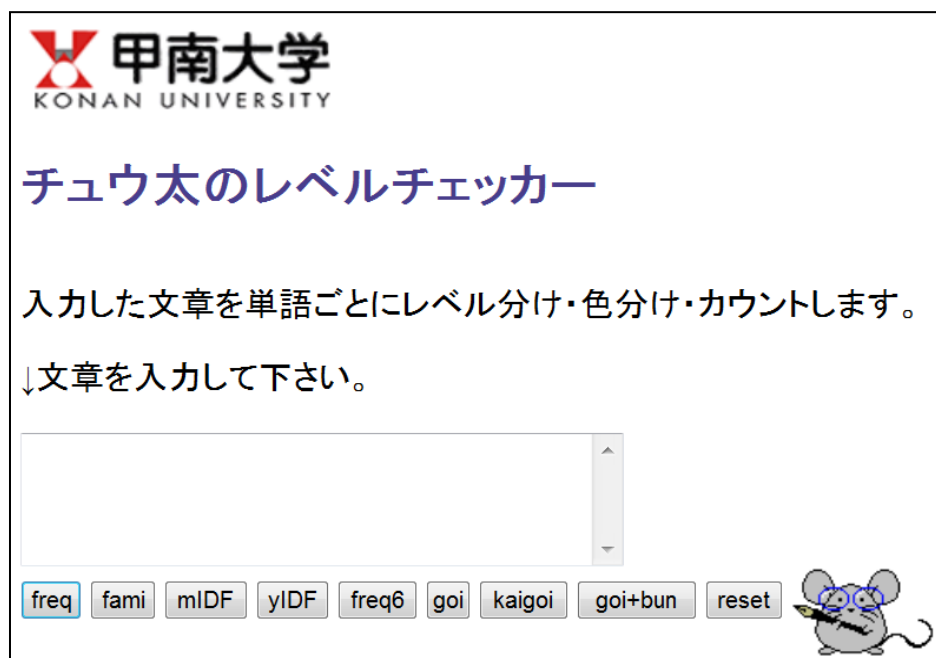
難易度判定システムでは、β版同様、入力された文章を形態素解析システム MeCab によって単語に区切り、文に含まれるすべての単語について出題基準に合わせてレベル判定を行う。ただし、固有名詞に関しては、MeCab によって固有名詞であるかどうかを判定して、一般の単語とは別に固有名詞数として扱うことにした。

一方、コンピュータによる自動処理を行うに際して、文の定義としては、次のような基準を設け、これをもとに文数をカウントすることにした。

- a. 文は、改行／句点の次の文字から次の改行／句点までとする
- b. 文字列が改行のみで区切られている場合は、文とみなさない
(題名、見出し等、文ではない要素を排除するため)
- c. b の場合は、文の平均文字数や単語数の算定にも加えない

難易度判定システムは、以上の基準をもとに、入力された文章の一文ごとの単語数をカウントする。図1が、難易度判定システムの入力画面である。

図1：難易度判定システムの入力画面



甲南大学
KONAN UNIVERSITY

チュウ太のレベルチェッカー

入力した文章を単語ごとにレベル分け・色分け・カウントします。
↓文章を入力して下さい。

freq fami mIDF yIDF freq6 goi kaigoi goi+bun reset

このシステムを使うには、図1の画面中央のテキストボックスに文章を入れ、**goi+bun**のボタンを押す。ここでは、『介護福祉士新カリキュラム学習ワークブック』の第一章を入力した。図2が結果画面である。

図2：難易度判定システムの結果画面

level6: Black, level5: SaddleBrown, level4: Purple, level3: Blue, level2: Green, level1: Magenta, others: Red

《1》「人間を理解する」とは、過去、現在の日常の生活状況における人間と、未来への希望といった志向性を含めた、生活の営みの歴史を理解することである。(73字)
 《2》人間の尊厳とは、人間が個人として尊重されることを意味する。(32字) 人々の生活の営みにおいて目指すべきものという意味で生活支援の基本的原理であり、「理念的価値」である。(50字)
 《3》自立とは、心身の障害等によって生活支障を生じている人々が日常生活をセルフケアによって営む生活の自立だけでなく、心理的に他者への依存から脱却し、自ら意思決定(自己決定)し、可能な限り、社会における何らかの役割をもち、活動することである。(120字)
 《4》かつて、自立とは専門職の視点からとらえられていたが、近年では、生活支援を利用する人(利用者)にこそ優位性があるとされている。(65字) 利用者の意思・責任といった主体性に基づいてニーズを表明し、生活支援の諸サービスを選定・利用してより良い生活を営むことを「生活自立」といい、介護福祉士等の支援を「自立支援」という。(89字)
 《5》自立の前提には、「自律」がある。(19字) 自立を現実的、効果的に達成するには、自分で立てた規範に従って、自分のことは自分でやっていくという精神的自律が求められる。(60字) また、精神的な自律と具体的な生活自立は、利用者の人格的態度によって統合され、不可分であるととらえられる。(52字)

一文の長さの平均: 62.2字

レベル: -0.2 N0(超級)

総数	語彙総数	level4	level3	level2	level1	その他	助詞,助動詞,接続詞	記号類
341	277	22	26	62	24	26	117	64
123.1%	100.0%	7.9%	9.4%	22.4%	8.7%	9.4%	42.2%	23.1%
(137)	(123)	(13)	(7)	(42)	(16)	(18)	(27)	(14)
111.4%	100.0%	10.6%	5.7%	34.1%	13.0%	14.6%	22.0%	11.4%

結果画面上の本文では、入力された文章を MeCab によって単語に区切り、各単語のレベルを出題基準に準じて次のように色分け表示している。

4 級：紫 3 級：青 2 級：緑

1 級：桃 その他（級外）：赤

助詞・助動詞・接続詞および記号類：黒

一文あたりの文字数の表示は、β版に準じて行う。文字数の色分け表示の基準は次のとおりである。

40 文字未満：黒 40 文字以上：青 60 文字以上：赤

また、文章全体の一文あたりの平均文字数も、表示する。

一方、画面中央にあるレベルの表示が、3.で得られた難易度判定式によって算出した数値とそれを基にしたレベル判定である。このレベル判定は、上述したように新日本語能力試験やCEFRを視野に入れて初級（N5）から超級（N0）までの6段階で評価している。図2においては、入力された文章の難易度の数値は、-0.2であり、超級（N0）と判定されている。

また、レベル表示の下にある表は出題基準を基にした級別の単語数と単語総数（語彙総数）に対する級ごとの語彙の比率を求めたものである。上の段が延べ語数、下の段が異なり語数である。

このシステムは、現在、<http://basil.is.konan-u.ac.jp/chuta/level/>で公開している。

6. おわりに

今回開発したシステムは、現在試験運用中のため上記サイトでのみ公開しているが、将来は、筆者らが開発し1999年からWeb上（<http://language.tiu.ac.jp/>）で無償公開している読解学習支援システム『リーディング・チュウ太』（川村2009）に組み入れる予定である。

今回得られた難易度判定式は、評価実験で有効性が確かめられたものの、あくまでも暫定的なものである。難易度判定用の学習データが異なれば、当然のことながら難易度判定式も異なってくる。また、単語の難易度判定基準に用いている出題基準が、学習者にとっての単語の難易度をどこまで忠実にあらわしているのかという問題も残っている。そのため、今後の課題として、より多くの難易度判定用学習データを用いた分析を続けるとともに、難易度判定結果が学習者の感じる難易度と一致しているかについて、学習者による判定を元にした実証実験を行う必要がある。また、出題基準とは異なったレベル別語彙リストを用いた分析も行っていく必要がある。さらに、非漢字圏学習者のためには、漢字の使用率等も考慮に入れた分析も行っていきたい。

一方、構文解析技術が進歩し、より精度の高い解析が行えるようになった際には、複文や名詞修飾節等の構文の複雑さを指標に加えた改良を行っていく予定である。

* 本稿はCAJLE2012での発表をもとにしている。本研究の一部は、科学研究費（平成23年度基盤（B）課題番号21320095および24年度基盤（B）課題番号24320096）によるものである。ここに記して謝意を表す。

参考文献

- 内田聖也・北村達也・川村よし子（2011）「文章難易度に寄与する構文の自動検出」『日本語教育学会春季大会予稿集』289-290 日本語教育学会
- 介護福祉士受験ワークブック編集委員会（2010）『介護福祉士新カリキュラム学習ワークブック』中央法規
- 川村よし子（1998）「読解のためのレベル判定システムの構築」『日本語教育方法研究会誌』Vol.5, No.2, 10-11 日本語教育方法研究会
- 川村よし子・北村達也（1998）「語彙チェッカーを用いた読解支援システム -読解学習支援環境の構築のために-」『日本教育工学会研究報告集』JET98-6, 29-34 日本教育工学会
- 川村よし子（2009）『チュウ太の虎の巻-日本語教育のためのインターネット活用術』くろしお出版
- 川村よし子（2011）「文章の難易度判定システム構築のための基礎調査」『ヨーロッパ日本語教育』15, 171-178 ヨーロッパ日本語教師会
- 川村よし子・野村愛・名藤杏子・金庭久美子・斉木美紀・武田亜寿香・西山陽子・橋本智・林奈実・藤田朋世・牧田弘子（2011）『介護福祉士新カリキュラム学習ワークブック』第3分冊 静岡県
- 川村よし子（2012）「インターネット上で利用可能な文章の難易度判定システムの開発」『ヨーロッパ日本語教育』Vol.16, 194-198 ヨーロッパ日本語教師会

工藤拓 (2006) MeCab: Yet Another Part-of-Speech and Morphological Analyzer, <http://mecab.sourceforge.net>

黒橋禎夫 (2000) 「けっこうやるな, KNP」 『情報処理』 41 (11)
1215-1220 情報処理学会

柴崎秀子 (2009) 「日本語リーダビリティ公式の構築と測定ツールの開発」 『特定領域研究日本語コーパス平成 20 年度公開ワークショップ研究成果報告会予稿集』 155-160

中村慶太・北村達也・川村よし子 (2012) 「検索エンジンを用いた主格省略文の自動判定」 『日本語教育方法研究会誌』 Vol.19, No1, 4-5 日本語教育方法研究会

水嶋博志・内田聖也・北村達也・川村よし子 (2011) 「学習者にとって難解な構文の自動検出」 『日本語教育方法研究会誌』 Vol.18, No.1, 64-65 日本語教育方法研究会

劉志宇・内田理 (2012) 「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」 『情報処理学会研究報告』 2011 号 1-5 情報処理学会

Sato, Satoshi, Matsuyoshi, Suguru, & Kondoh, Yohsuke. (2008). *Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus*. LREC-08.