https://doi.org/10.31634/cjs.2023.59.0251
Comparative Japanese Studies
2023. 12, Vol.59, pp.251-273

日本語のテキストの可読性公式の高度化研究*

– 旧JLPTの語彙と文法の過去問の導入 –

金 曙 泳**

Abstract — In this paper, we used the text of the former JLPT's reading comprehension questions as a sample and added a data survey on the new JLPT's vocabulary and grammar questions. Through this, I verified various factors that are thought to affect the difficulty of the text and calculated a readability formula with more meaningful independent variables than the existing readability formula.

Specifically, a database of reading comprehension texts from the past 20 years of the former JLPT and vocabulary and grammar exam questions from the new JLPT from the past 10 years was constructed to conduct basic analysis. I then performed Pierce correlation analysis and multiple linear regression analysis on the readability model, which allowed us to select meaningful independent variables and calculate the readability formula by assigning weights to each variable. Lastly, the significance of the readability formula was confirmed through verification of the 'Variance expansion factor; VIF' for the readability formula.

[Japanese text Readability Level Formula] - AR2: 0.7923, p-value: < 2.2e-16

$$y = 4.3028887 - 0.0415297x_1 - 0.0003788x_2 - 0.0139127x_3 - 2.1304148x_4 - 1.2913411x_5 - 2.1453816x_6 - 2.4320030x_7 - 3.7829578x_8 - 0.5638206x_9 - 1.3076988x_{10}$$

- y : Japanese text level = Readability
- x_1: Number of sentences per paragraph, x_2: Number of morphemes per text, x_3: Number of morphemes per sentence, x_4: Chinese character ratio, x_5: N4 grade Kanji ratio, x_6: N3 grade Kanji ratio, x_7: N2 grade Kanji ratio, x_8: N1 grade Kanji ratio, x_9: N4 grade vocabulary ratio, x_{10}: N1 grade vocabulary ratio

Key words : Readability, Readability formula, JLPT, Difficulty level of Text, Japanese-Readability

1. はじめに

テキストの可読性(Readability・以下, 可読性)を簡単に定義すると「テキストを読ん

* 이 논문은 2022년도 동덕여자대학교 연구비 지원에 의하여 수행된 것임(연구번호: 202205957).
This study was supported by the Dongduk Women's University grant(No. 202205957)

** 同徳女子大学校 日語日本学科 副教授

で理解できる程度」或いは「テキストの読みやすさ」などになる[1]。このような可読性は多様な分野において意味のある尺度であって，教育分野のみならず，テキストが使われるあらゆる場面と関連付けて考えられる。例えば言語教育においては学習者のレベルに合わせて表記(漢字・仮名など)・語彙・文法の数と難易度，また句や文の長さなどをコントロールしなければならない。また小学生向けのマニュアル，ネイティブではない外国人向けの案内書など，特定の読者が想定された情報テキストにも可読性は考慮されなければならない。

本稿では，このような可読性に影響を及ぼすと思われる因子を統計的な方法を持って計算することによって日本語のテキストにおける可読性を数値化できる「日本語のテキストの可読性公式(Japanese Readability formula・以下，可読性公式」[2]を高度化することを目的とした。そうすることによってより精度の高い可読性公式が提案できると思われる。

2. 先行研究と問題提議

2.1. 可読性の因子

1節では可読性に影響を及ぼすと思われる因子を言及したが，佐藤理史他(2012:1−2)はその因子を次の(1)と(2)のように「テキスト側の要因」と「読み手側の要因」に区分して説明した。しかし，(2)の読み手側の場合，読み手の個人差などを考慮すると一律で考えられない上で，定量的に数値化するのも難しい。よって本稿では「読み手の要因」を固定的なものとし，テキスト側の要因の中で数値化できる因子のみを抽出して可読性公式を算出することにする。

(1)［読むもの(テキスト) 側の要因］

a. テキストの表示の見やすさ・見にくさ(legibility)：表示媒体、フォーマット、印刷状態など

b. テキスト表現のやさしさ・難しさ(readability)：文章表現、語彙、文体など

1)『日本国語大辞典』第二版(2010), Oxford Advanced Learner's Dictionary. 8th Edition(2010).

2) 可読性テストまたは数値的可読性指標とも呼ばれる

c. 書かれている内容の複雑さ

(2) [読み手側の要因]

a. その言語の運用能力

b. 書かれている内容に対する背景知識

c. その時の身体状況 佐藤理史他(2012:1-2)

2.2. 可読性公式

可読性公式に関する研究は1920年代から始まった以来, 多くの研究が行われてきたが,「テキスト側の要因」を主な因子としてテキストの可読性を数値化できる公式を算出しようとした研究が最も盛んに行われてきた。1928年のVogel and Washburneによる研究以来, 50年代まで約31件の公式が発表されてきたがその中でもっとも予測性が高い公式とされるのがDale and Chall(1948)で, 最も頻繁に言及される公式はFlesch(1948)である。その詳しい公式は以下の(3)のようになる。

(3) [代表的な可読性公式の例]

a. Dale and Chall(1948)

$$X_{c50} = 0.1579x_1 + 0.0496x_2 + 3.6365$$

- x_1: Dale listの3,000単語に含まれていない単語の比率(%)
- x_2: 単語数で計算した文章の長さ

b. Flesch(1948)

$$R.E. = 206.835 - 0.846wl - 1.015sl$$

- wl: 100単語当たりの音節数の平均 - sl: 1文章当たりの単語数の平均

一方, 日本語における可読性公式に関する研究は以下の(4)aの浅野陽子他(1991), (4)bの川村よし子他(2013), (4)cの金曙泳(2015)などがあげられる。まず(4)aの浅野陽子他(1991)は小学校から高校までの国語の教科書に掲載された20件のテキストで見られる「平仮名・カタカナ」と「句点」の数を因子として可読性公式を算出した。続いて(4)bの川村よし子他(2013)は旧日本語能力試験(以下, 旧JLPT)の出題基準にあわせた「語彙レベル」と「文章あたり語彙数」という二つの因子を独立変数と設

定し, 重回帰分析(Multiple regression analysis)を行って可読性公式を算出した。最後に(4)cの金(2015)は旧JLPTの読解の過去問(20年間)にピアソンの相関分析(Pearson correlation analysis)と線形回帰分析(Linear regression analysis)などの統計分析を行い,「文章あたり形態素」「漢字の割合」「N4レベル漢字」「N3レベル漢字」「N2レベル漢字」「N1レベル漢字」が主な因子だということを明らかにした。その結果(4)cのような可読性公式を算出した。

(4)［日本語のテキストの可読性公式］

a. PGV = −0.17*ph − 0.28*pk − 3.49*pe + 27.92
   1) PGV:学年レベルを基準とした可読性の測度
      ph:平仮名の出現頻度 pk:カタカナの出現頻度
      pe:句点の出現頻度 浅野陽子他(1991:1576)

b. レベルN＝−3.020350224 −0.108713131×[a] + 5.9903188×[b]
      + 5.3699195×[c] + 1.0666679×[d] + 9.7980957×[e]
   1) a:一文あたりの平均単語数, b:総N1／総単語数, c:総N2／総単語数, d:総N3／総単語数, e:総N4／総単語数(b〜eにおいて記号・固有名詞を除く)
   2) [a]と[e]のみがp＜0.01 川村よし子他(2013:23−24)

c. 日本語のテキストの可読性レベル公式 − AR2: 0.7848, p−value: < 2.2e−16

$$y = 4.041029 - 0.011292x_1 - 0.022071x_2 - 0.016339x_3 - 0.025853x_4 - 0.026349x_5 - 0.046882x_6$$

   1) y: 日本語のテキストレベル
   2) x_1: 文章当たり形態素, x_2: 漢字比率(%), x_3: N4漢字数, x_4: N3漢字数, x_5: N2漢字数, x_6: N1漢字数 金曦泳(2015:37)

(4)aの浅野(1991)は比較的に早い時期における可読性公式ではあるが, 調査対象と因子が比較的に少ない。また, (2)bの川村(2013)は統計分析における詳細な内容が公開されていない上, 統計的に有意味ではない因子の相関係数([b], [c], [d])を可読性公式に反映するなど信頼に乏しい点が見られる。その一方で(4)c金(2015)の日本語のテキストの可読性公式は大規模のデータベースを基盤として統計的に有意味な独立変数のみを因子として取り入れた可読性公式である。しかし, 管見の限り, 金(2015)を含めて先行研究における可読性公式に新JLPTの語彙と文法における出題基準を因子として統計的な分析を行った研究は見られない。従って本稿では(4)cの可

読性公式を基盤としてJLPTの語彙と文法の出題基準を加えて有意味な独立変数を見いだすことによってより高度な可読性公式の算出することを目指した。

3. 研究方法

3.1. 可読性公式のための分析項目と研究対象

可読性公式を算出するための統計分析を行う前に，本稿では調査対象である旧JLPTの過去問の「難易度(級・レベル，1〜4)」に影響を及ぼす，つまり可読性に影響を及ぼすと思われる因子として以下の(5)のような項目に，新たに(6)の因子を取り入れた。

(5) [旧JLPTの過去問のデータベースにおける分析項目及び目的]

a. 1段落当たりの文数：段落の構造とその簡単さを把握
b. 文章・段落・文当たりの形態素数：言葉の長さとその簡単さを把握
c. 漢字・平仮名・カタカナの比率：「全体の文字における漢字の比率」と「仮名と漢字の比率」を通じて難易度を把握　金曙泳(2014:217)

(6) [可読性公式の高度化のための追加した分析項目]

a. 全体の文字数
b. 全体漢字における「N1・N2・N3・N4の漢字」，それぞれの比率
c. 全体語彙における「N1・N2・N3・N4の語彙」，それぞれの比率
d. 全体文法項目における「N1・N2・N3・N4の文法項目」，それぞれの比率

続いて本稿では，過去20年間の旧JLPT(1〜4)における文法・読解の過去問のテキストのみのデータベース[3]を標本として構築し(旧JLPTテキストDB)，新JLPT(N1〜N5)の語彙・文法の過去問に基づいて新JLPTの漢字・語彙・文法の辞書データベース(新JLPT辞書DB)[4]も構築した。そこで「旧JLPTテキストDB」における(5)のような基本的な数値をまとめた上で，「新JLPT辞書DB」に基づいて(6)のような数値を算出し

3) 20年間(1990年〜2009年)の旧日本語能力試験(JLPT)の全ての過去問
4) 10年間(2010年〜2020年)の新日本語能力試験(JLPT)の過去問の語彙と文法項目

たが, そのローデータ(n=437)の一部は以下の絵1のようになる[5]。

Lv	Moji	Sen_Para	Morp_Txt	Morp_Para	Morp_Sen	Kanji	Kana_Kanji	N4_Kanji	N3_Kanji	N2_Kanji	N1_Kanji	N4_Goi	N3_Goi	N2_Goi	N1_Goi	N4_Bun	N3_Bun	N2_Bun	N1_Bun
1	902	5	573	114.6	22.92	0.19	3.75	0.29	0.15	0.21	0.05	0.24	0.14	0.05	0.03	0.88	0.08	0.03	0.01
1	1014	3.2	613	122.6	38.31	0.25	2.7	0.26	0.23	0.15	0.08	0.27	0.16	0.09	0.06	0.83	0.11	0.04	0.03
1	44	1	27	27	27	0.18	4	0.38	0.12	0	0	0.17	0	0	0	0.83	0.08	0.08	0
1	80	1	44	44	44	0.26	2.67	0.33	0.33	0.1	0.05	0.32	0.09	0.05	0.05	0.71	0.12	0.12	0.06
1	67	1	38	38	38	0.16	4.64	0.36	0.27	0.18	0.09	0.33	0.13	0.07	0.13	0.76	0.19	0.05	0
1	61	1	34	34	34	0.18	4.27	0.09	0.18	0.09	0	0.31	0.19	0	0	0.86	0.07	0.07	0
1	133	3	78	78	26	0.24	2.66	0.53	0.38	0.03	0	0.31	0.17	0.06	0.23	0.45	0.45	0.1	0
1	1119	3.3	685	68.5	20.76	0.24	2.74	0.39	0.18	0.18	0.1	0.35	0.11	0.16	0.06	0.86	0.08	0.05	0.01
1	880	1.82	559	50.82	27.95	0.28	2.23	0.33	0.14	0.15	0.07	0.4	0.15	0.05	0.01	0.88	0.08	0.03	0.01
1	143	5	93	93	18.6	0.21	3.2	0.37	0.13	0.1	0.13	0.38	0.16	0.06	0.06	0.87	0.1	0.03	0

<図 1> 旧JLPTの標本に対する分析項目の統計

3.2. 統計分析

本稿における統計分析には主に「R」[6]を利用したが, 詳しくは可読性に影響を及ぼすと思われる因子に対してピアソンの相関分析(Pearson correlation analysis)を通じて相関関係が認められる変数を設定した。この結果をもってさらに重回帰分析(Multiple regression analysis)を行い, 有意味な独立変数とそれぞれに重みを設定し、統計的に有意味な回帰モデル, つまりより高度な可読性公式が算出した。

4. 日本語のテキストのレベルと可読性

4.1. 相関関係 - ピアソンの相関分析(Pearson correlation analysis)

本論ではまず, 日本語のテキストの可読性と関連していると思われる(5)と(6)のような因子が「旧JLPT過去問の難易度」, つまり可読性と有意義な相関関係であるかを確認するために, ピアソンの相関分析を行った。その因子別の結果をまとめるとは以下の表1のようになる

5) http://japanese.or.kr/japaneseutill/Readability%20Tool%20Series/AJ-JpnRa_Tool_Thesis.aspx

6) 「R」は統計とデータ解析に特化したオープンソース・プログラミング言語。- http://www.r-project.org 参照

<表 1>「可読性」と「因子」におけるピアソンの相関分析

因子	相関係数(cor)	有意確率(p-value)
文字数	-0.4013329	p-value < 2.2e-16
1段落当たりの文数	-0.2480754	p-value = 1.493e-07
1文章当たりの形態素数	-0.4011881	p-value < 2.2e-16
1段落当たりの形態素数	-0.5194431	p-value < 2.2e-16
1文当たりの形態素数	-0.5912492	p-value < 2.2e-16
全体における漢字の比率	-0.75941	p-value < 2.2e-16
仮名対漢字の比率	0.493658	p-value < 2.2e-16
N4漢字の比率	-0.3619967	p-value = 5.619e-15
N3漢字の比率	-0.7419955	p-value < 2.2e-16
N2漢字の比率	-0.6765688	p-value < 2.2e-16
N1漢字の比率	-0.5556342	p-value < 2.2e-16
N4語彙の比率	-0.3375239	p-value = 4.186e-13
N3語彙の比率	-0.6053932	p-value < 2.2e-16
N2語彙の比率	-0.5469646	p-value < 2.2e-16
N1語彙の比率	-0.5215667	p-value < 2.2e-16
N4文法の比率	0.4686711	p-value < 2.2e-16
N3文法の比率	-0.3959253	p-value < 2.2e-16
N2文法の比率	-0.2807968	p-value = 2.319e-09
N1文法の比率	-0.09789028	p-value = 0.04081

(7) [「N3漢字の比率」におけるPearson's product-moment correlationの例]

data: jlpt$Lv and jlpt$N3_Kanji
t = -23.084, df = 435, **p-value < 2.2e-16**
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.7814118 -0.6966817
sample estimates: **cor -0.7419955**

表1の因子における「相関係数(Correlation Coefficient, rco, r)」と「有意確率(p-value)」によると, 本稿で可読性に影響を及ぼすと仮定した全ての因子はその相関性において偏差はあるものの, 統計的に有意味であることを確認できた。例えば, (7)の「N3漢字の比率」の場合, 相関係数の絶対値(|-0.7419955|)が「1」に近い[7]上で, 有

7) a. 0 ≤ r ≤ 0.2 : 低い相関関係　　b. 0.2 < r ≤ 0.4 : 少し低い相関関係
c. 0.4 < r ≤ 0.7 : 少し高い相関関係　　d. 0.7 < r ≤ 0.9 : 高い相関関係

意確率も「0.001」より小さい($p < 0.001$)[8]ので，高い相関関係が認められる。

4.2. 線形回帰分析(Linear Regression Analysis)

前節における個別の因子と可読性との相関分析を基盤として，本節では全ての因子を「独立変数」に，因子に影響される可読性を「従属変数」に設定して，線形回帰分析を行ったがその結果は以下の表2のようになる。

<表 2> 線形回帰分析の結果

```
Call : lm(formula = Lv ~ ., data = jlpt)
Residuals:
     Min       1Q   Median       3Q      Max
-1.67579 -0.24225  0.01014  0.29363  1.18493
Coefficients:  Estimate  Std.  Error  t value  Pr(>|t|)
(Intercept)   1.0784056  4.2190962   0.256  0.798385
Moji          0.0035059  0.0009801   3.577  0.000388 ***
Sen_Para     -0.0826839  0.0333538  -2.479  0.013570 *
Morp_Txt     -0.0061140  0.0015977  -3.827  0.000150 ***
Morp_Para     0.0015416  0.0013804   1.117  0.264738
Morp_Sen     -0.0180320  0.0042317  -4.261  2.52e-05 ***
Kanji        -1.7382077  0.4033801  -4.309  2.05e-05 ***
Kana_Kanji    0.0029825  0.0043703   0.682  0.495345
N4_Kanji     -1.1823047  0.2047996  -5.773  1.52e-08 ***
N3_Kanji     -1.6847894  0.3198095  -5.268  2.21e-07 ***
N2_Kanji     -2.0630543  0.4270596  -4.831  1.91e-06 ***
N1_Kanji     -3.3779881  0.6310111  -5.353  1.43e-07 ***
N4_Goi       -0.7903032  0.3003468  -2.631  0.008821 **
N3_Goi       -0.8430990  0.4399874  -1.916  0.056024 .
N2_Goi       -0.7933212  0.4506471  -1.760  0.079072 .
N1_Goi       -1.6776407  0.6825115  -2.458  0.014376 *
N4_Bun        3.3553862  4.2170662   0.796  0.426678
N3_Bun        3.1505309  4.2225446   0.746  0.456014
N2_Bun        2.3232790  4.2629571   0.545  0.586050
N1_Bun        2.3186106  4.3142080   0.537  0.591253
```

e. $0.9 < r \leq 1.0$: より高い相関関係或いは完全な相関関係

8) 例えば，「$p < 0.001$」の場合，仮説が間違える確率(有意性，可能性，probability)は0.1%以下，つまり少なくとも99.9%ぐらい信頼できる。「$p < 0.05$」の場合は95%ぐらい信頼できることになる。

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.4287 on 417 degrees of freedom
Multiple R-squared:  0.8121,    Adjusted R-squared:  0.8035
F-statistic: 94.86 on 19 and 417 DF,  p-value: < 2.2e-16
```

表2の結果を解釈すると, 有意確率が「0.0001」より小さい(p-value: < 2.2e-16)ので, 有意水準「0.05」でこの回帰モデルは統計的に妥当である。また, この回帰モデルには有意ではない独立変数[9]があるため, 「Multiple R-squared」ではなく「Adjusted R-squared」によって分析結果を解釈すると, 今回の回帰モデルの説明率は80.35%にまで上って, 有意味な回帰モデルであることが確認できる。ところが, 「1段落たりの形態素数:Morp_Para」のように帰無仮説である独立変数も存在するため, より高度な可読性公式を算出するためには独立変数を選択する必要があるが, 詳しくは次節で述べることにする。

4.3. 重回帰分析(Multiple regression analysis)

本節では, 独立変数の選択とその重みを確認するために「R」のパッケージのAIC(Akaike information criterion, Akaike:1974)の重回帰分析の変数選択方法を採択して分析を行った。そうすることによって, 最も統計的に有意味な回帰モデル, つまりより高度な可読性公式が算出することを目指した。変数選択方法には三つの方法があるが, ①変数増加法 (Forward Selection)による結果は表3で, ②変数減少法(Backward Elimination)による結果は表4, 最後に③変数増減法(forward-backward stepwise selection method)による結果は表5のようになる。

<表 3> ①AIC基盤の変数増加法(Forward Selection)の結果

```
Start:  AIC = -720.84
Lv ~ Moji + Sen_Para + Morp_Txt + Morp_Para + Morp_Sen + Kanji + Kana_Kanji + N4_Kanji
+ N3_Kanji + N2_Kanji + N1_Kanji + N4_Goi + N3_Goi + N2_Goi + N1_Goi + N4_Bun + N3_Bun
+ N2_Bun + N1_Bun
Call: lm(formula = Lv ~ Moji + Sen_Para + Morp_Txt + Morp_Para + Morp_Sen + Kanji +
```

9) 例えば独立変数の中で「1段落たりの形態素数」の「Morp_Para」はp-valueが「0.005 < 0.264738」であって有意ではない。つまり, 帰無仮説。

Kana_Kanji + N4_Kanji + N3_Kanji + N2_Kanji + N1_Kanji + N4_Goi + N3_Goi + N2_Goi + N1_Goi + N4_Bun + N3_Bun + N2_Bun + N1_Bun, data = jlpt)

Coefficients:

(Intercept)	Moji	Sen_Para	Morp_Txt	Morp_Para	Morp_Sen	Kanji
Kana_Kanji	N4_Kanji	N3_Kanji	N2_Kanji	N1_Kanji	N4_Goi	N3_Goi
N2_Goi	N1_Goi	N4_Bun	N3_Bun	N2_Bun	N1_Bun	
1.078406	0.003506	−0.082684	−0.006114	0.001542	−0.018032	−1.738208
0.002982	−1.182305	−1.684789	−2.063054	−3.377988	−0.790303	−0.843099
−0.793321	−1.677641	3.355386	3.150531	2.323279	2.318611	

Residuals:

Min	1Q	Median	3Q	Max
−1.67579	−0.24225	0.01014	0.29363	1.18493

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.0784056	4.2190962	0.256	0.798385	
Moji	0.0035059	0.0009801	3.577	0.000388	***
Sen_Para	−0.0826839	0.0333538	−2.479	0.013570	*
Morp_Txt	−0.0061140	0.0015977	−3.827	0.000150	***
Morp_Para	0.0015416	0.0013804	1.117	0.264738	
Morp_Sen	−0.0180320	0.0042317	−4.261	2.52e−05	***
Kanji	−1.7382077	0.4033801	−4.309	2.05e−05	***
Kana_Kanji	0.0029825	0.0043703	0.682	0.495345	
N4_Kanji	−1.1823047	0.2047996	−5.773	1.52e−08	***
N3_Kanji	−1.6847894	0.3198095	−5.268	2.21e−07	***
N2_Kanji	−2.0630543	0.4270596	−4.831	1.91e−06	***
N1_Kanji	−3.3779881	0.6310111	−5.353	1.43e−07	***
N4_Goi	−0.7903032	0.3003468	−2.631	0.008821	**
N3_Goi	−0.8430990	0.4399874	−1.916	0.056024	.
N2_Goi	−0.7933212	0.4506471	−1.760	0.079072	.
N1_Goi	−1.6776407	0.6825115	−2.458	0.014376	*
N4_Bun	3.3553862	4.2170662	0.796	0.426678	
N3_Bun	3.1505309	4.2225446	0.746	0.456014	
N2_Bun	2.3232790	4.2629571	0.545	0.586050	
N1_Bun	2.3186106	4.3142080	0.537	0.591253	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4287 on 417 degrees of freedom

Multiple R−squared: 0.8121, Adjusted R−squared: 0.8035

F−statistic: 94.86 on 19 and 417 DF, p−value: < 2.2e−16

<表 4> ②AIC基盤の変数減少法(Backward Elimination)の結果

Step: **AIC = −727.39**

Lv ~ Moji + Sen_Para + Morp_Txt + Morp_Sen + Kanji + N4_Kanji + N3_Kanji + N2_Kanji + N1_Kanji + N4_Goi + N3_Goi + N2_Goi + N1_Goi

	Df	Sum of Sq	RSS	AIC
<none>			77.584	−727.39
− N2_Goi	1	0.5151	78.099	−726.49
− N3_Goi	1	0.6752	78.259	−725.60
− N1_Goi	1	1.1488	78.733	−722.96
− N4_Goi	1	1.4536	79.037	−721.27
− Moji	1	2.4713	80.055	−715.68
− Sen_Para	1	2.7148	80.299	−714.36
− Morp_Txt	1	2.7911	80.375	−713.94
− Kanji	1	4.2508	81.835	−706.08
− N2_Kanji	1	4.7268	82.311	−703.54
− Morp_Sen	1	5.1339	82.718	−701.39
− N1_Kanji	1	5.7128	83.297	−698.34
− N3_Kanji	1	5.9496	83.533	−697.10
− N4_Kanji	1	6.8042	84.388	−692.65

Call:

lm(formula = Lv ~ Moji + Sen_Para + Morp_Txt + Morp_Sen + Kanji + N4_Kanji + N3_Kanji + N2_Kanji + N1_Kanji + N4_Goi + N3_Goi + N2_Goi + N1_Goi, data = jlpt)

Coefficients:

(Intercept)	Moji	Sen_Para	Morp_Txt	Morp_Sen	Kanji	N4_Kanji
N3_Kanji	N2_Kanji	N1_Kanji	N4_Goi	N3_Goi	N2_Goi	N1_Goi
4.368168	0.003576	−0.049300	−0.006195	−0.014863	−1.793753	−1.224384
−1.785028	−2.132750	−3.494265	−0.822311	−0.839525	−0.748322	−1.683533

Residuals:

Min	1Q	Median	3Q	Max
−1.76551	−0.25345	0.03122	0.29614	1.22362

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.3681685	0.0928474	47.047	< 2e−16	***
Moji	0.0035757	0.0009741	3.671	0.000273	***
Sen_Para	−0.0493002	0.0128143	−3.847	0.000138	***
Morp_Txt	−0.0061948	0.0015880	−3.901	0.000111	***
Morp_Sen	−0.0148626	0.0028092	−5.291	1.96e−07	***

Kanji	−1.7937530	0.3726013	−4.814	2.06e−06 ***
N4_Kanji	−1.2243841	0.2010227	−6.091	2.53e−09 ***
N3_Kanji	−1.7850276	0.3134132	−5.695	2.31e−08 ***
N2_Kanji	−2.1327497	0.4201206	−5.077	5.77e−07 ***
N1_Kanji	−3.4942651	0.6261024	−5.581	4.28e−08 ***
N4_Goi	−0.8223109	0.2920982	−2.815	0.005103 **
N3_Goi	**−0.8395248**	**0.4375577**	**−1.919**	**0.055700 .**
N2_Goi	**−0.7483224**	**0.4465399**	**−1.676**	**0.094512 .**
N1_Goi	−1.6835331	0.6726760	−2.503	0.012700 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.4283 on 423 degrees of freedom
Multiple R−squared: 0.8097, Adjusted R−squared: 0.8039
F−statistic: 138.5 on 13 and 423 DF, p−value: < 2.2e−16

<表 5> ③AIC基盤の変数増減法(forward−backward stepwise selection method)の結果

Step: **AIC = −727.39**
Lv ~ Moji + Sen_Para + Morp_Txt + Morp_Sen + Kanji + N4_Kanji + N3_Kanji + N2_Kanji + N1_Kanji + N4_Goi + N3_Goi + N2_Goi + N1_Goi + N2_Bun

	Df	Sum of Sq	RSS	AIC
<none>			77.230	−727.39
− N2_Bun	1	0.3543	77.584	−727.39
+ Morp_Para	1	0.2650	76.965	−726.89
− N2_Goi	1	0.4998	77.729	−726.57
+ N1_Bun	1	0.1090	77.121	−726.00
+ N4_Bun	1	0.0944	77.135	−725.92
+ Kana_Kanji	1	0.0872	77.142	−725.88
+ N3_Bun	1	0.0400	77.190	−725.61
− N3_Goi	1	0.6971	77.927	−725.46
− N1_Goi	1	1.0352	78.265	−723.57
− N4_Goi	1	1.4514	78.681	−721.25
− Moji	1	2.4678	79.697	−715.64
− Sen_Para	1	2.5226	79.752	−715.34
− Morp_Txt	1	2.7931	80.023	−713.86
− Kanji	1	4.2470	81.477	−705.99
− N2_Kanji	1	4.8848	82.114	−702.59
− Morp_Sen	1	5.0184	82.248	−701.88
− N3_Kanji	1	5.2764	82.506	−700.51

```
- N1_Kanji   1      5.7327   82.962    -698.10
- N4_Kanji   1      6.5974   83.827    -693.56

Call:
lm(formula = Lv ~ Moji + Sen_Para + Morp_Txt + Morp_Sen + Kanji + N4_Kanji + N3_Kanji
+ N2_Kanji + N1_Kanji + N4_Goi + N3_Goi + N2_Goi + N1_Goi + N2_Bun, data = jlpt)
Coefficients:
(Intercept)        Moji    Sen_Para    Morp_Txt    Morp_Sen       Kanji    N4_Kanji
   N3_Kanji    N2_Kanji    N1_Kanji     N4_Goi      N3_Goi      N2_Goi      N1_Goi
    N2_Bun
   4.375799    0.003573   -0.047711   -0.006197   -0.014706   -1.792966   -1.207769
  -1.707455   -2.173335   -3.500418   -0.821695   -0.853248   -0.737225   -1.603882
  -1.023632
Residuals:
     Min       1Q   Median       3Q      Max
 -1.7128   -0.2462   0.0284   0.2994   1.2227
Coefficients:
             Estimate  Std.  Error   t value   Pr(>|t|)
(Intercept)   4.3757986  0.0929068   47.099   < 2e-16 ***
Moji          0.0035732  0.0009731    3.672  0.000271 ***
Sen_Para     -0.0477114  0.0128510   -3.713  0.000233 ***
Morp_Txt     -0.0061970  0.0015863   -3.907  0.000109 ***
Morp_Sen     -0.0147063  0.0028084   -5.237  2.59e-07 ***
Kanji        -1.7929656  0.3721901   -4.817  2.03e-06 ***
N4_Kanji     -1.2077687  0.2011554   -6.004  4.15e-09 ***
N3_Kanji     -1.7074549  0.3179924   -5.369  1.31e-07 ***
N2_Kanji     -2.1733347  0.4206690   -5.166  3.69e-07 ***
N1_Kanji     -3.5004177  0.6254265   -5.597  3.94e-08 ***
N4_Goi       -0.8216951  0.2917759   -2.816  0.005088 **
N3_Goi       -0.8532485  0.4371857   -1.952  0.051637 .
N2_Goi       -0.7372246  0.4461180   -1.653  0.099170 .
N1_Goi       -1.6038818  0.6743670   -2.378  0.017834 *
N2_Bun       -1.0236325  0.7356817   -1.391  0.164835
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.4278 on 422 degrees of freedom
Multiple R-squared:  0.8106,    Adjusted R-squared:  0.8043
F-statistic:   129 on 14 and 422 DF,  p-value: < 2.2e-16
```

続いて変数選択方法による結果の中で, 同じく最も小さいAICの値(-727.39)を持つ②「表4・変数減少法の回帰モデル」と③「表5・変数増減法の回帰モデル」に的を

絞って両方における比較を行った。その結果から表5の③「変数増減法の回帰モデル」を採択することにしたが(表6, 514.7656:**514.7654**), その中には独立変数「N3_Goi:N3語彙の比率」と「N2_Bun:N2文法の比率」の有意確率が0.05より大きく(0.051637と0.164835), 統計的に有意味ではない。よって, 最終的な回帰モデル, つまり可読性公式においてはこれらの独立変数を排除したが, その結果が以下の(8)bと表7のようになる。

<表 6> ②「変数減少法による回帰モデル」と③「変数増減法によるモデル」の比較

	df	AIC
jlpt_lm_ba	15	514.7656
jlpt_lm_bo	16	**514.7654**

(8)［重回帰分析による回帰モデルの独立変数の設定］

a. 変数

1) 従属変数: 可読性

2) 独立変数: 文字数, 1段落当たりの文数, 1文章当たりの形態素数, 1段落たりの形態素数, 1文当たりの形態素数, 全体における漢字の比率, 仮名対漢字の比率, N4漢字の比率, N3漢字の比率, N2漢字の比率, N1漢字の比率, N4語彙の比率, N3語彙の比率。N2語彙の比率, N1語彙の比率, N4文法の比率, N3文法の比率, N2文法の比率, N1文法の比率

b. AIC分析結果―回帰モデル・可読性公式

[文字数 + 1段落当たりの文数 + 1文章当たりの形態素数 + 1文当たりの形態素数 + 全体における漢字の比率 + N4漢字の比率 + N3漢字の比率 + N2漢字の比率 + N1漢字の比率 + N4語彙の比率 + N1語彙の比率]

<表 7> AIC分析結果―回帰モデル・可読性公式

```
Call:  lm(formula = Lv ~ Moji + Sen_Para + Morp_Txt + Morp_Sen + Kanji +
    N4_Kanji + N3_Kanji + N2_Kanji + N1_Kanji + N4_Goi + N1_Goi,
    data = jlpt)
Residuals:
     Min       1Q   Median       3Q      Max
-1.64097 -0.25517  0.02416  0.30198  1.20023
```

```
Coefficients:
             Estimate  Std.  Error      t value     Pr(>|t|)
(Intercept)   4.324109     0.091383     47.318    < 2e-16 ***
Moji          0.003352     0.000972      3.448    0.000620 ***
Sen_Para     -0.049766     0.012868     -3.867    0.000127 ***
Morp_Txt     -0.005830     0.001585     -3.679    0.000264 ***
Morp_Sen     -0.015465     0.002805     -5.513    6.12e-08 ***
Kanji        -2.032771     0.359962     -5.647    2.99e-08 ***
N4_Kanji     -1.283869     0.199065     -6.450    3.06e-10 ***
N3_Kanji     -2.146142     0.271780     -7.897    2.47e-14 ***
N2_Kanji     -2.494794     0.378220     -6.596    1.26e-10 ***
N1_Kanji     -3.591926     0.626265     -5.735    1.85e-08 ***
N4_Goi       -0.556422     0.271121     -2.052    0.040752 *
N1_Goi       -1.414810     0.665929     -2.125    0.034200 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.4301 on 425 degrees of freedom
Multiple R-squared:  0.8072,    Adjusted R-squared:  0.8022
F-statistic: 161.7 on 11 and 425 DF,  p-value: < 2.2e-16
```

```
Call:lm(formula = Lv ~ Moji + Sen_Para + Morp_Txt + Morp_Sen + Kanji +
    N4_Kanji + N3_Kanji + N2_Kanji + N1_Kanji + N4_Goi + N1_Goi,
    data = jlpt)
Coefficients:
(Intercept)         Moji     Sen_Para     Morp_Txt     Morp_Sen
   4.324109     0.003352    -0.049766    -0.005830    -0.015465
      Kanji     N4_Kanji     N3_Kanji     N2_Kanji     N1_Kanji
  -2.032771    -1.283869    -2.146142    -2.494794    -3.591926
     N4_Goi       N1_Goi
  -0.556422    -1.414810
```

以上の回帰モデルは「Adjusted R-squared: 0.8022」であって, 使われた独立変数が従属変数である可読性の変動を約80%ぐらい説明できるという意味になる。そしてp-valueが0.001より小さいので(p-value: < 2.2e-16), この回帰モデルが統計的に有意味であることも意味する。このような分析を通じて, 以下の(9)のような公式を算出できたが, これが日本語のテキストの可読性を計算できる公式になる。

(9) [日本語のテキストの可読性公式]

$$y = \alpha_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11}$$

$$y = 4.324109 + 0.003352x_1 - 0.049766x_2 - 0.005830x_3 - 0.015465x_4 - 2.032771x_5 - 1.283869x_6 - 2.146142x_7 - 2.494794x_8 - 3.591926x_9 - 0.556422x_{10} - 1.414810x_{11}$$

- y: 日本語のテキストの可読性
- α: Y切片, β_1～β_{11}:独立変数の回帰係数
- x_1:文字数, x_2:1段落当たりの文数, x_3:1文章当たりの形態素数, x_4:1文当たりの形態素数, x_5:全体における漢字の比率, x_6:N4漢字の比率, x_7:N3漢字の比率, x_8:N2漢字の比率, x_9:N1漢字の比率, x_{10}:N4語彙の比率, x_{11}:N1語彙の比率

4.4. 可読性公式に関する検証

本節では可読性公式を算出した重回帰分析における検証を行うことにするが, 詳しくは独立変数の間における線形従属関係の有無を確認することである。これを多重共線性(Multicollinearity)の問題というが, 独立変数と独立変数の間に相関関係が高い, つまり多重共線性が強いと回帰分析の結果に歪曲が生じる可能性が高くなるからである。従って本稿ではVIF(Variance Inflation Factor・分散拡大係数)により, 上記の(9)の公式に対して多重共線性の検証を行ったが, その結果は以下の表8のようになる。

<表 8> VIF(Variance Inflation Factor)による可読性公式の検証

Moji	Sen_Para	**Morp_Txt**	Morp_Sen	Kanji	N4_Kanji	N3_Kanji
246.181241	1.293181	**246.142835**	1.813057	2.596333	1.446117	2.275362
N2_Kanji	N1_Kanji	N4_Goi	N1_Goi			
1.950683	1.544081	1.434412	1.589157			

VIF値が「10」より大きいと多重共線性が存在すると考えられるが, 上記の表(8)をみると独立変数の中で「文字数:Moji」と「1文章当たりの形態素数:Morp_Txt」に多重共線性が存在することが分かる。従って, 本稿では①「文字数:Moji」の変数のみを排除, ②「1文章当たりの形態素数:Morp_Txt」の変数のみを排除, ③「文字数:Moji」と「1文章当たりの形態素数:Morp_Txt」, 両方を排除するという三つの場合を想定し

て, それぞれの回帰分析を行ったが, その結果が次の表9のようになる。

<表 9> VIF検証による独立変数の選択

①

```
Call: lm(formula = Lv ~ Sen_Para + Morp_Txt + Morp_Sen + Kanji + N4_Kanji +
    N3_Kanji + N2_Kanji + N1_Kanji + N4_Goi + N1_Goi, data = jlpt)
Residuals:
     Min       1Q   Median       3Q      Max
-1.6480  -0.2685   0.0452   0.2874   1.2579
Coefficients:
             Estimate Std.  Error     t value  Pr(>|t|)
(Intercept)  4.3028887   0.0923340   46.601   < 2e-16 ***
Sen_Para    -0.0415297   0.0128048   -3.243   0.001275 **
Morp_Txt    -0.0003788   0.0001132   -3.345   0.000895 ***
Morp_Sen    -0.0139127   0.0028037   -4.962   1.01e-06 ***
Kanji       -2.1304148   0.3634045   -5.862   9.15e-09 ***
N4_Kanji    -1.2913411   0.2015813   -6.406   3.96e-10 ***
N3_Kanji    -2.1453816   0.2752316   -7.795   4.99e-14 ***
N2_Kanji    -2.4320030   0.3825799   -6.357   5.31e-10 ***
N1_Kanji    -3.7829578   0.6317342   -5.988   4.51e-09 ***
N4_Goi      -0.5638206   0.2745563   -2.054   0.040627 *
N1_Goi      -1.3076988   0.6736543   -1.941   0.052893 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.4356 on 426 degrees of freedom
Multiple R-squared:  0.8018,    Adjusted R-squared:  0.7971
F-statistic: 172.3 on 10 and 426 DF,  p-value: < 2.2e-16
```

②

```
Call:lm(formula = Lv ~ Moji + Sen_Para + Morp_Sen + Kanji + N4_Kanji +
    N3_Kanji + N2_Kanji + N1_Kanji + N4_Goi + N1_Goi, data = jlpt)
Residuals:
      Min        1Q    Median        3Q       Max
-1.64251 -0.26959  0.04538  0.28590  1.26337
Coefficients:
             Estimate Std.  Error     t value  Pr(>|t|)
(Intercept)  4.299e+00   9.246e-02   46.496   < 2e-16 ***
Moji        -2.151e-04   6.958e-05   -3.091   0.00213 **
Sen_Para    -4.075e-02   1.282e-02   -3.180   0.00158 **
Morp_Sen    -1.385e-02   2.811e-03   -4.926   1.20e-06 ***
Kanji       -2.137e+00   3.641e-01   -5.870   8.79e-09 ***
N4_Kanji    -1.299e+00   2.019e-01   -6.433   3.37e-10 ***
N3_Kanji    -2.154e+00   2.757e-01   -7.813   4.39e-14 ***
N2_Kanji    -2.432e+00   3.834e-01   -6.344   5.74e-10 ***
N1_Kanji    -3.834e+00   6.319e-01   -6.067   2.89e-09 ***
```

<table>
<tr><td></td><td>
N4_Goi −5.634e−01 2.751e−01 −2.048 0.04115 *
N1_Goi −1.288e+00 6.748e−01 −1.909 0.05693 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.4364 on 426 degrees of freedom
Multiple R−squared: 0.801, <u>Adjusted R−squared: 0.7964</u>
F−statistic: 171.5 on 10 and 426 DF, p−value: < 2.2e−16
</td></tr>
<tr><td>③</td><td>
Call:lm(formula = Lv ~ Sen_Para + Morp_Sen + Kanji + N4_Kanji + N3_Kanji +
N2_Kanji + N1_Kanji + N4_Goi + N1_Goi, data = jlpt)
Residuals:
Min 1Q Median 3Q Max
−1.57030 −0.27127 0.04328 0.30126 1.28066
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.272069 0.092963 45.954 < 2e−16 ***
Sen_Para −0.038308 0.012920 −2.965 0.0032 **
Morp_Sen −0.014334 0.002834 −5.058 6.31e−07 ***
Kanji −2.135512 0.367712 −5.808 1.24e−08 ***
N4_Kanji −1.385175 0.201988 −6.858 2.46e−11 ***
N3_Kanji −2.264709 0.276148 −8.201 2.82e−15 ***
N2_Kanji −2.482474 0.386817 −6.418 3.68e−10 ***
N1_Kanji −4.283149 0.621067 −6.896 1.93e−11 ***
N4_Goi −0.552465 0.277792 −1.989 0.0474 *
N1_Goi −1.142016 0.679801 −1.680 0.0937 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.4408 on 427 degrees of freedom
Multiple R−squared: 0.7966, <u>Adjusted R−squared: 0.7923</u>
F−statistic: 185.8 on 9 and 427 DF, p−value: < 2.2e−16
</td></tr>
<tr><td>AIC</td><td>
df AIC
① jlpt_lm_bo_last_moji 12 <u>526.6603</u>
② jlpt_lm_bo_last_mT 12 528.2999
③ jlpt_lm_bo_last_moji_mT 11 535.9924
</td></tr>
<tr><td>③のVIF</td><td>
Sen_Para Morp_Txt Morp_Sen Kanji N4_Kanji N3_Kanji
1.248624 1.225157 1.766387 2.580267 1.445946 2.275360
N2_Kanji N1_Kanji N4_Goi N1_Goi
1.946162 1.531999 1.434323 1.585700
</td></tr>
</table>

表9における分析を通じて, ①の回帰モデルが最も有意味であることが確認できた(「説明率(0.7971)」と「AICの値:526.6603」, ①>②>③)。また, 表9の「③のVIF」の

ように，①の回帰モデルには多重共線性の問題もなかった。このような分析結果を反映した最終的な可読性公式は以下の(10)のようになる。

(10) [日本語のテキストの可読性公式]

$$y = \alpha_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10}$$

$$y = 4.3028887 - 0.0415297 x_1 - 0.0003788 x_2 - 0.0139127 x_3 - 2.1304148 x_4 - 1.2913411 x_5 - 2.1453816 x_6 - 2.4320030 x_7 - 3.7829578 x_8 - 0.5638206 x_9 - 1.3076988 x_{10}$$

- y: 日本語のテキストの可読性
- α: Y切片，β_1～β_{11}: 独立変数の回帰係数
- x_1: 1段落当たりの文数，x_2: 1文章当たりの形態素数，x_3: 1文当たりの形態素数，x_4: 全体における漢字の比率，x_5: N4漢字の比率，x_6: N3漢字の比率，x_7: N2漢字の比率，x_8: N1漢字の比率，x_9: N4語彙の比率，x_{10}: N1語彙の比率

5. まとめ

本稿では旧JLPTの読解の過去問のテキストを標本として，新JLPTまで調査範囲を広めて新JLPTの語彙と文法の過去問など，テキストの難易度に影響を及ぼすと思われる因子を検証することによって以下の(11)のような有意味な独立変数を持つより高度な可読性公式を算出することができた。

(11) [日本語のテキストの可読性公式] － AR^2: 0.7923, p-value: < 2.2e-16

$$y = 4.3028887 - 0.0415297 x_1 - 0.0003788 x_2 - 0.0139127 x_3 - 2.1304148 x_4 - 1.2913411 x_5 - 2.1453816 x_6 - 2.4320030 x_7 - 3.7829578 x_8 - 0.5638206 x_9 - 1.3076988 x_{10}$$

- y: 日本語のテキストの可読性
- x_1: 1段落当たりの文数，x_2: 1文章当たりの形態素数，x_3: 1文当たりの形態素数，x_4: 全体における漢字の比率，x_5: N4漢字の比率，x_6: N3漢字の比率，x_7: N2漢字の比率，x_8: N1漢字の比率，x_9: N4語彙の比率，x_{10}: N1語彙の比率

しかし，可読性には2節で述べたように読み手の領域はもちろん，JLPTのような語学試験の評価基準以外にも様々な因子が影響を及ぼすと思われる。従って稿者は今後の課題として，読み手による新たな因子を見出す方法を考慮する一方，旧JLPTの読解の過去問のテキストのみならず新JLPTの読解の過去問における資料調査を行うことにして，より大規模の標本に基づいてより高度な可読性公式を算出したいと思う。また，このような可読性公式を日本語教育など，実際的に活用して行きたい。

〈参考文献〉

浅野陽子・小川克彦(1991)「日本文の可読性の測度と表示速度への応用」,『情報処理学会論文誌』32(12), 一般社団法人情報処理学会, pp.1574−1582.

阿辺川武・八木豊・戸次徳久・澤谷孝志・奥村学・仁科喜久子・杉本茂樹, 傅亮(2003)「日本語学習システム「あすなろ」開発の新しい展開―構文学習とその評価―」,『情報処理学会第65回大会論文集』3T2−6, 情報処理学会, pp.1−4.

稲積宏誠(2012)「基本文型理解のためのICTを活用した日本語学習支援システムの開発 ―プロトタイプシステムの概要―」,『第九回日本語教育・日本研究シンポジウム論文集』, 香港日本語教育研究会, pp.1−8.

乾裕子・岡田直之(2000)「長い文は常にわかりにくいか? : わかりにくさの要因とその依存関係」,『情報処理学会研究報告』2000(11), 一般社団法人情報処理学会, pp.63−70.

影山功・宮崎佳典・長谷川由美(2009)「Readability式を用いたオンライン外国語学習環境の構築」,『情報科学技術フォーラム講演論文集』8(4), FIT(電子情報通信学会・情報処理学会)運営委員会, pp.481−484.

川村よし子(1998)「読解のためのレベル判定システムの構築: 語彙チェッカーの開発と活用」,『日本語教育方法研究会誌』5−2, 日本語教育方法研究会, pp.10−11.

川村よし子・北村達也(2013)「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」,『Journal CAJLE』Vol.14, Canadian Association for Japanese Language Education, pp.18−30.

北尾謙治・北尾S. キャスリーン(2011)「Readability and vocabulary level of reading passages in Japanese University entrance exams」,『文化情報学』6(1), 同志社大学, pp.11−20.

北村雅則・石川美紀子・加藤良徳・棚橋尚子・山口昌也(2009)「作文支援システムTEachOtherSの運用と成果分析」,『名古屋学院大学論集 言語・文化篇』21(1), 名古屋学院大学総合研究所, pp.43−54.

金曙泳(2014)「日本語のテキストの可読性分析―舊JLPTの過去問を基盤とした日本語のテキストの漢字及び漢字語彙の可読性の判断プログラムの開発―」,『日本言語文化』27号, 韓国日本言語文化学会, pp.357−382.

_____(2015)「日本語のテキストの可読性レベル分析―旧日本語能力試験の過去問のデータベースに対する統計的な検証を基盤として―」,『日本学報』Vol.103, 韓国日本学会, pp.21−40.

清川英男(1978)「リーダビリティ研究の概観」,『淑徳大学研究紀要』12, 淑徳大学, pp.65−82.

小島健輔・佐藤理史(2009)「現代日本語書き言葉均衡コーパスに対する難易度付与(テキスト評価とリーダ

ビリティ)」,『Technical report of IEICE. Thought and language』109(84), The Institute of Electronics, Information and Communication Engineers, pp.175−184.
小島健輔・佐藤理史・藤田篤(2009)「文字bigramモデルを用いた日本語テキストの難易度推定」,『言語処理学会第15回年次大会発表論文集』, 言語処理学会, pp.897−900.
小林雄一郎・北尾 謙治(2010)「Comparing graded readers and authorized English language textbooks in junior and senior high schools : from the viewpoint of vocabulary and readability」,『文化情報学』5(1), 同志社大学, pp.1−14.
佐藤理史(2008a)「日本語テキストの難易度を測る(特集 言語処理研究の新展開―計算機と言語学の対話に向けて)」,『月刊言語』37(8), 大修館書店, pp.54−57.
_____(2008b)「日本語テキストの難易度判定ツール『帯』」,『Japio YEAR BOOK 2008 寄稿集』一般財団法人日本特許情報機構, pp.52−57.
_____(2011)「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』52(4), pp.1777−1789.
_____(2013)「テキストの難易度と語の分布」,『情報処理学会研究報告』2013−NL−213(6), 一般社団法人情報処理学会, pp.1−11.
_____(2014)「コンピュータの国語力を大学入試問題で測る(産業日本語関連)」,『Japio year book』日本特許情報機構, pp.274−277.
佐藤理史・柏野和佳子(2012)「テキストの難易度に対する人間の判断と機械の判断」,『第1回 コーパス日本語学ワークショップ予稿集』, 国立国語研究所, pp.195−202.
柴崎秀子・玉岡賀津雄(2010)「国語科教科書を基にした小・中学校の文章難易学年判定式の構築」,『日本教育工学会論文誌』33(4), 日本教育工学会, pp.449−458.
祖国威・加納敏行(2010)「構文的な分かりやすさを評価する可読性評価技術」,『言語処理学会 第16回年次大会 発表論文集』, 言語処理学会, pp.1082−1085.
高木裕子(1991)「速読用読解教材開発に向けて―リーダビリティー研究を基礎にして」,『関西外国語大学留学生別科日本語教育論集』Vol.1, pp.66−85.
建石由佳・小野芳彦・山田尚勇(1988)「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, 情報処理学会, pp.1−8.
田中健二(1996)「新たなリーダビリティ測定法: 試案」,『JACET全国大会要綱』35, 一般社団法人大学英語教育学会, pp.347−348.
土屋武久(1998)「リーダビリティの決定要因についての一考察 : 大学リーディング教材の計量的分析から」,『JACET全国大会要綱』37, 一般社団法人大学英語教育学会, pp.50−51.
日本国語大辞典編集委員会編(2001)『日本国語大辞典』第二版, 小学館, Edition staff of Nihon Kokugo Daijiten(2001)『the Nihon Kokugo Daijiten』2nd edition, pp.1−1469.
樋口晶彦・樋口高子(2012)「A Study of Automated L2 Writing Evaluation by Japanese College Students : Eva Text Analysis」,『鹿児島大学教育学部研究紀要 教育科学編』64, 鹿児島大学, pp.1−9.
平本哲嗣(1994)「テクストの読み易さに関与する要因に関する一考察」,『中国地区英語教育学会研究紀要』(24), 中国地区英語教育学会, pp.107−114.
寒川クリスティーナ・フメリャク(2014)「日本語教科書コーパスの構築と分析 : 日本語学習者のためのリーダビリティー測定に向けて」,『日本語教育方法研究会誌』19(2), 日本語教育方法研究会, pp.4−5.
黄永熙・金曙泳(2014)「オンライン大学の日本語基礎課程の漢字コンテンツ−日本語能力試験漢字および常用漢字との対照−」,『日本語学研究』40号, 韓国日本語学会, pp.165−180.
ホン・ジョンハ他(2011)「テキストの水準と可読性:韓国語の学習教材を用いた検証と応用」,『言語情報』Vol.12, 高麗大学言語情報研究所, pp.111−148.
萬戸克憲(2000)「テキストの難易の測定とリーディング指導」,『英米評論』15, 桃山学院大学, pp.91−115.

本岡直子(1987)「Readabilityに関する一考察：特にスキーマ理論の立場から」,『中国地区英語教育学会研究紀要』(17), 中国地区英語教育学会, pp.1-8.
山田純・バーノンC.(1977)「Readabilityに関する一調査」,『中国地区英語教育学会研究紀要 : CASELE research bulletin』(7), 中国地区英語教育学会, pp.79-84.
Akaike, H. (1974), "A new look at the statistical model identification", IEEE Transactions on Automatic Control 19(6), IEEE, pp.716-723.
Cohen, J.(1988), *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*, New York: Academic Press, pp.1-590.
Dale, E. & Chall, J. S.(1948), "A formula for predicting readability." *Educational Research Bulletin 27,* Taylor & Francis, Ltd., pp.37-54.
Dale, E. & Chall, J. S.(1956), "Developing Readable Materials." In Henry, N. B. (ed.) *Adult Reading*, The 55th Yearbook of the NSSE, Part II, University of Chicago, pp.218-244.
Flesch, R.(1946), *The art of plain talk*. New York: Harpers, pp.1-223.
Flesch, R.(1948), "A New Readability Yardstick." *Journal of Applied Psychology 32*, pp.221-233.
Gilliland, J.(1972), *Readability*, University of London Press, pp.1-128.
KAWAMURA Yoshiko(2013), Implementation and Evaluation of a System for Determining Japanese Text-Level Difficulty, *Journal CAJLE*, pp.18-30.
KOJIMA Kensuke and SATO Satoshi(2009), Readability Assignment to Balanced Corpus of Contemporary Written Japanese. *IEICE technical report 109(84),* IEICE, pp.13-18.
Klare, G. R.(1963), *The measurement of readability*. T. Ames, IA:Iowa State University Press, pp.1-358.
Satoshi Sato, Suguru Matsuyoshi, Yohsuke Kondoh(2008), Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), pp.654-660.
Oxford University Press(2010), *Oxford Advanced Learner's Dictionary*. Oxford University Press, pp.1-1820.
Spache, G.(1953), "A new readability formula for primary grade reading materials", *Elementary School Journal* 53(7), The University of Chicago, pp.410-413.

논문투고일 : 2023.11.10.
논문심사일 : 2023.11.24.
심사확정일 : 2023.12.11.

필자 인적사항
성명 : (한글) 김유영, (한자) 金曘泳, (영어) Kim Yu Young
소속 : 동덕여자대학교 일어일본학과
논문영문제목 : Research on the Advancement of Readability Formula for Japanese Texts: Addition of Factors for former JLPT Vocabulary and Grammar Past Questions
E-mail : yuiyu@hotmail.com

<국문요지>

본고에서는 구JLPT 독해 기출문제의 텍스트를 표본으로, 새롭게 신JLPT의 어휘와 문법 기출문제에 대한 자료조사를 추가하여 텍스트의 난이도에 영향을 미칠 것이라 생각되는 다양한 인자들 검증하는 것을 통해 기존의 가독성 공식보다 더욱 유의미한 독립변수를 가진 가독성 공식을 산출했다.

구체적으로는 구JLPT의 과거 20년간의 독해 지문과 신JLPT의 어휘와 문법 기출문제에 대한 데이터베이스 구축을 통해 표본 텍스트에 대한 기초 분석을 수행했으며, 이를 기반으로 피어스 상관분석과 다중 선형회귀 분석을 통해 유의미한 독립변수 선택 후 가중치를 부여하여 가독성 공식을 산출하고 이를 최종적으로 분산 팽창 인수 검증을 통해 유의성을 확인했다.

[일본어 텍스트 가독성 공식] – AR2 : 0.7923, p-value: < 2.2e-16

$$y = 4.3028887 - 0.0415297x_1 - 0.0003788x_2 - 0.0139127x_3 - 2.1304148x_4 - 1.2913411x_5 - 2.1453816x_6 - 2.4320030x_7 - 3.7829578x_8 - 0.5638206x_9 - 1.3076988x_{10}$$

- y : 일본어 텍스트 가독성
- x_1 : 단락 당 문장 수, x_2 : 텍스트 당 형태소 수, x_3 : 문장 당 형태소 수,
  x_4 : 한자비율, x_5 : N4 출제기준 한자 비율, x_6 : N3 출제기준 한자 비율,
  x_7 : N2 출제기준 한자 비율, x_8 : N1 출제기준 한자 비율,
  x_9 : N4 출제기준 어휘 비율, x_{10} : N1 출제기준 어휘 비율

주제어 : 가독성, 가독성 공식, JLPT, 난이도, 일본어 가독성