일본어 텍스트의 가독성 레벨 분석

— 구舊일본어능력시험 기출문제 데이터에 대한 통계적 검증을 기반으로 —

김 유 영*
yuiyu1004@dongduk.ac.kr

<ABSTRACT>

Text readability is a statistical measure of the level and the difficulty of the text, and is influenced by various factors. This work aimed to extract only those factors that can be quantified in a Japanese text from among the various factors that may affect the readability of the text, which were then used to analyze the level through statistical techniques. Through these processes I verified the validity of each of the factors, and have developed a formula that forms the core of the readability level analytical system.

Specifically, by using Pierce correlation analysis on the 'JLPT reading section Corpus', which consists of 437 reading texts extracted from the last 20 years of previous JLPT, I propose that 'Morpheme number per sentence', 'Kanji percentage', 'N4 grade Kanji percentage', 'N3 grade Kanji percentage', 'N2 grade Kanji percentage', 'N1 grade Kanji percentage' are applicable readability metrics for Japanese text. Using the above readability metrics, the readability was calculated by performing a multiple linear regression analysis using the following Japanese text readability formula.

[Japanese text Readability Level Formula] - AR^2: 0.7848, p-value: < 2.2e-16

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

- y : Japanese text level = Readability
- x_1: Morpheme number per sentence, x_2: Kanji percentage, x_3: N4 grade Kanji percentage, x_4: N3 grade Kanji percentage, x_5: N2 grade Kanji percentage, x_6: N1 grade Kanji percentage

Key words: Readability, Readability Formula, Readability metrics for Japanese text, Corpus, Text Level, JLPT, Correlation Analysis, Linear Regression Analysis, Readability Analyzer, AJ-JpnRa Tool

1. 들어가며

사전적 정의에 따르면, 텍스트 가독성이란 텍스트를 '읽고 이해할 수 있는 정도' 혹은 텍스트를 '읽기 쉬운 정도'라고 간단히 정리할 수 있는데[1], 최근에는 텍스트의 수준 및 난이도를 평가하는 통계적 척도로도 정의 및 사용되는 경우가 많다. 이러한 텍스트 가독성은 다양한 요인에 의해 영향을 받는데, 다음의 (1)과 같이 '리더(Reader・読者, 이하 리더) 영역'과 '텍스트(Text・テキスト, 이하 텍스트) 영역'이라는 크게 두 차원으로 구분하여 생각해 볼 수 있다. 佐藤理史 외(2012:1-2)에 따르면 아래 (1)a의 텍스트 영역을 1)텍스트의 보기 좋음 정도 legibility(매체, 양식, 인쇄 상태 등)와 2)텍스트 표현의 난이도readability(문장표현, 어휘, 문체

* 동덕여자대학교 조교수, 일본어학

1) 『日本国語大辞典』第2版(2010), *Oxford Advanced Learner's Dictionary*. 8th Edition(2010).

등) 그리고 3)내용의 복잡성 등으로 설명하기도 했다.

(1) [텍스트 가독성에 영향을 미치는 요인]
a. 리더 영역 : 사전 지식, 독해 능력, 흥미 및 동기
b. 텍스트 영역 : 내용, 언어형식(길이・표기・어휘・문법・문체 등), 구성형식(디자인・구성・폰트 등) 김유영(2014:2)

이러한 텍스트 가독성은 다양한 차원에서 의미 있는 척도로서, 예를 들어 언어 교육에서 어느 정도 가독성 레벨을 가진 교재를 채택하여 학습자에게 제공하는 것이 효과적일 것인가? 특정 정보제공을 목적으로 하는 텍스트가 과연 목표로서 상정하고 있는 독자의 수준에 적합한 가독성 레벨을 가지고 있는가? 그리고 평상시와 다른 재난 상황과 같은 급박한 상황에는 어느 정도의 가독성 레벨의 텍스트가 적당한가? 등, 언어교육뿐만 아니라 텍스트와 연관된 일상생활의 실질적인 문제에 이르기까지 많은 연관성을 갖고 있는 척도가 바로 텍스트 가독성이다.

본 연구는 위와 같은 일본어 텍스트의 가독성 레벨에 영향을 미치리라 생각되는 다양한 요인들을 통계적 기법을 통해 분석하는 것을 통해 각각의 요인들의 유효성을 검증하고, 이를 바탕으로 한발 더 나아가 텍스트 가독성 레벨 판정 시스템의 핵심이 되는 가독성 레벨 공식을 개발하는 데에 목적이 있다.

2. 선행연구 및 문제점

텍스트 가독성에 관한 연구는 1920년대 초부터 미국에서 활발히 진행되어 왔는데, 清川英男(1978:1-2)은 이들 가독성 관련 연구의 관점을 다음의 (2)와 같이 크게 네 가지로 분류했다.

(2) a. 어휘 : 이어(異語), 추상어, 다음절어 등의 백분율. 특정 어휘목록에 없는 어휘의 비율 등을 가독성 지표로 설정
b. 문장 구조 : 단문의 수, 문장의 길이 등을 가독성 지표로 설정
c. 클로즈 법(Cloze procedure) : 일정 간격을 두고 빈칸을 설정한 텍스트를 피험자가 올바르게 메꿀 비율을 가독성 지표로 설정
d. 가독성 공식 : 어휘와 문장 구조 등 복합적인 요인을 조합한 공식을 통해 가독성 예측

Klare(1963)에 의하면, 그중에서도 (2)d와 같이 텍스트의 가독성을 산출할 수 있는 공식을 개발하고자 하는 연구가 가장 활발하게 진행되어 왔으며, 1928년의 Vogel and Washburne 이래 50년대까지 약 31개의 공식이 발표되었다고 한다. 그중에서 가장 예측성이 높은 공식이 Dale and Chall(1948), 그리고 가장 빈번하게 사용된 공식이 Flesch(1948)라고 하는데, 두 공식

을 간단히 정리하면 다음의 (3)과 같다.

(3) a. Dale and Chall(1948)

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

- x_1: Dale list의 3,000단어에 포함되어 있지 않은 단어의 백분율
- x_2: 단어의 수로 계산한 문장의 길이

b. Flesch(1948)

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

- wl: 100 단어의 모든 음절 수
- sl: 한 문장 내 단어 수의 평균

위 공식 이외에도 많은 텍스트 가독성 공식이 계속해서 발표되었으나, Dale(1956)이 언급한 바와 같이, 가독성에는 공식만으로는 판별할 수 없는 요소가 존재하는 것이 사실이다. 예를 들어 길이가 짧고 문장구조 또한 단순한 문장이라고 하더라도 그 내용이 심오하여 가독성이 떨어지는 경우(Conceptual difficulty)가 있을 수 있으며, 문장의 개수와 길이 등 수치가 동일한 두 개의 텍스트가 있다고 하더라도 각각의 텍스트가 어떻게 구성되어 있는가에 따라 독자가 느끼는 가독성이 달라질 수 있다(Organization). 또한 텍스트의 내용 별 독자의 흥미여하에 따라(Content) 그리고 텍스트의 매체 및 인쇄상태 등에 따라서도(Format) 가독성이 달라질 수 있다. 마찬가지로 平本哲嗣(1994)는 텍스트의 수치 이외에도 독자의 의미규칙과 관련된 지식, 담화에 관한 지식, 일반적인 세계지식 등이 가독성에 영향을 미친다고 언급했으며, 本岡直子(1987)는 특정 텍스트와 그 텍스트를 무작위로 섞어 재배열한 텍스트를 가독성 공식을 통해서 분석하면 서로 동일한 결과를 얻을 수 있기 때문에, 가독성 공식을 통해서는 문레벨은 가능할 수 있을지라도 텍스트 전체의 가독성을 파악하는 것은 어렵다고 언급했다.

하지만, 우선 平本(1994:112-113)가 스스로 밝히고 있는 바와 같이, 일관성 있는 텍스트 가독성 분석을 위해서는, 담화라는 관점에서 가독성을 논하기 전에 담화 분석의 타당성 향상이 우선되어야 할 것이다. 그리고 무엇보다 가독성 공식을 통해 텍스트 가독성 레벨을 도출했다고 하더라도, 그것이 텍스트 가독성의 모든 범주를 가독성 공식만으로 설명했다는 것을 의미하지는 않는다. 즉, 本岡(1987)와 平本(1994)가 언급한 바와 같은 요인들도 텍스트 가독성에 영향을 미치는 요인임에 분명하지만, 선행연구들의 가독성 공식에 반영된 요인들 또한 텍스트 가독성에 영향을 미치는 요인이라고 볼 수 있다. 따라서 앞서 (1)에서 언급한 바와 같이 광의의 텍스트 가독성은 '가독성 공식'과 '그 외의 요인들'을 종합하여 설명할 수 있을 것으로 생각된다. 이에 본고에서는 (1)a'리더 영역'이 고정되어 있다고 가정하고, 텍스트 영역에서 수치화 할 수 있는 요인들만을 추출하여 가독성 공식에 반영하여, 협의의 텍스트 가독성 레벨 공식을 개발하고자 한다.

한편, 이와 같은 일본어 텍스트의 가독성 공식에 관한 선행연구로는 우선 佐藤理史(2008)

등을 들 수 있는데, ‘言語表現の難しさ-構成要素の難しさ(単語や文字の難しさ)’를 텍스트 가독성의 요인이라 상정하고, 구체적으로는 문중의 한자의 난이도와 비율 그리고 학년별 단어의 사용을 기준으로 가독성 분석을 실시했다. 하지만 일본어를 모국어로 하는 화자를 독자로 상정한 까닭에 일본어 텍스트 가독성 레벨에 대한 구분 기준을 비모국어 화자에게는 적용하기 어려운 부분이 있다.

한편 川村よし子 외(2013) 등의 경우 일본어 텍스트의 어휘를 일본어능력시험 출제기준에 대입한 수치(구舊일본어능력시험 1～4급)와 ‘문장 당 단어 수’의 수치를 독립변수로 설정하고 다중 선형회귀 분석을 실시하여 일본어 텍스트 가독성 공식을 도출했다. 그러나 川村(2013) 등은 통계분석의 상세한 내용을 명확히 제시하고 있지 않기 때문에, 각각의 급수(구舊일본어능력시험 1～4급)의 어휘가 전체 텍스트 난이도에 얼마나 영향을 미치는지 확인하기 어렵다. 그리고 무엇보다 가독성 레벨로 설정한 N0～N5의 분석 대상 텍스트 데이터의 분류 기준이 모호하고 자의적이다. 또한 통계분석 결과를 밝히지는 않았으나 유의하지 않다고 언급한 ‘구JLPT 4급의 단어 비율’의 상관계수를 가독성 공식에 반영하는 등, 통계분석 결과를 신뢰하기 어렵다.

3. 연구방법

이에 본고에서는 지금까지의 가독성 연구를 기반으로 다음 같은 연구방법을 통해 효과적이고도 객관적인 텍스트 가독성 판단의 기준을 검토하고자 한다. 우선, 지금까지 실시된 구舊일본어능력시험(이하, 구JLPT)의 독해 지문을 코퍼스 데이터베이스화 하여, 이를 기반으로 다양한 통계 기법을 사용하여 일본어 ‘텍스트 가독성’과 상관관계에 있다고 판단되는 ‘텍스트 분석항목’을 도출하고, 최종적으로는 일본어 텍스트 가독성 공식을 개발하고자 한다.

3.1. 연구 대상 및 가독성 분석을 위한 텍스트 영역 분석항목

본고에서는 김유영(2014:215-217)의 기준에 따라 최근 20년간[2]의 구舊JLPT(1급～4급) 독해・문법 영역 중 독해 지문만을 선별하여 ‘JLPT 독해 지문 데이터베이스’를 구축하고, 이 데이터베이스를 다음의 (4)와 같은 항목을 기준으로 통계분석을 실시했다. 그리고 각각의 구舊JLPT 지문은 해당 레벨(1～4급)을 그대로 수치화 했다(e.g. 구JLPT 3급 지문의 ‘가독성 레벨’ 수치 = 3).

(4) [JLPT 독해 지문 데이터베이스 분석 항목 및 목적]
a. 단락 당 문장 수 : 단락 구조와 단조로움의 정도를 파악
b. 텍스트・단락・문장 당 형태소 수 : 텍스트・단락・문장의 길이 및 이를 통해 텍스트의 단조

2) 최근 20년간(1990년~2009년), 구舊일본어능력시험(JLPT) 기출문제.

로움의 정도를 파악

c. 한자・히라가나・가타카나 비율 : '전제 문자 대비 한자 비율'과 '가나 VS 한자의 사용 비율'을 통해 난이도 정도를 파악 김유영(2014:217)

위 (4)의 구체적인 데이터(n = 437)는 그림1과 같으며, 다음의 웹페이지를 통해 모든 수치를 확인가능하다.

http://japanese.or.kr/japaneseutill/ReadabilityTool%20Series/AJ-JpnRa_Tool_Thesis.aspx

2. 일본어 텍스트 가독성 분석용 통계자료 - 2015.03.03

2.1 '텍스트 분석항목' 통계자료

No	가독성 레벨	단락당 문장	텍스트당 형태소	단락당 형태소	문장당 형태소	한자 비율	가나대 한자비	N5한자	N4한자	N3한자	N2한자	N1한자
1	1	2.25	318	79.5	35.33	43.53%	0.99	17.12%	33.33%	27.48%	15.32%	6.76%
2	1	4	150	150	37.5	38.34%	0.74	17.53%	17.53%	41.24%	12.37%	11.34%
3	1	6	140	140	23.33	33.19%	0.55	20.51%	28.21%	23%	20.51%	7.69%
4	1	1.5	176	88	58.67	32.38%	0.56	23.08%	34.07%	26.37%	15.38%	1%
5	1	6	157	157	26.17	24.82%	0.36	30.88%	16.18%	16.18%	26.47%	10.29%
6	1	2.33	506	56.22	24.1	24.05%	0.35	26.73%	21.78%	36.14%	10.89%	4.46%
7	1	3.5	457	114.25	32.64	29.53%	0.48	21.36%	13.64%	34.09%	26.36%	4.55%
8	1	3.25	517	129.25	39.77	35.53%	0.65	40.55%	23.37%	21.31%	11.68%	3.09%
9	1	4.14	843	120.43	29.07	30.70%	0.5	26.18%	28.77%	27.59%	10.85%	6.60%
10	1	3	213	106.5	35.5	33.44%	0.56	19.27%	33.03%	26%	11.93%	10.09%
11	1	3	103	103	34.33	29.19%	0.48	23.40%	21.28%	21.28%	23.40%	10.64%
12	1	11	179	179	16.18	19.66%	0.27	26.32%	40.35%	14.04%	14.04%	5.26%

그림 1. 일본어 텍스트 가독성 분석을 위한 '텍스트 분석항목' 통계자료(일부 발췌)

3.2. 통계 분석

구JLPT의 급수별 독해지문의 경우, 급수가 낮아질수록 절대적 텍스트양이 줄어들기 때문에, 단순 양적 비교는 큰 의미를 갖지 못한다. 따라서 본고에서는 통계분석을 실시하고자 하는데, 간단한 분석은 엑셀로, 복잡한 통계 처리를 요하는 경우에는 통계 프로그램인 'R 통계 패키지[3]'를 이용했다. 우선 일본어 텍스트 가독성과 텍스트 분석항목간의 피어슨 상관분석을 통해 상관성이 인정되는 변수를 설정하고, 이를 기반으로 다중 선형회귀분석을 실시하여 통계적으로 유의미한 회귀모형, 즉 가독성 공식을 도출했다.

4. 일본어 텍스트 수준과 가독성

4.1 상관관계 - 피어슨 상관분석

우선 일본어 텍스트 가독성 분석을 위해 텍스트 가독성과 연관관계에 있다고 가정한 '텍스트 분석항목'((4)와 그림1 참조) 각각의 값이 가독성 레벨과 통계적으로 유의미한 상관성을

3) 통계 계산과 그래픽을 위한 프로그래밍 언어이자 소프트웨어. - http://www.r-project.org 참조.

갖는지 여부를 확인하기 위해 피어슨 상관분석을 수행했으며, 그 결과는 다음의 표1과 그림2·3과 같다.

표 1. '가독성 레벨'과 '텍스트 분석항목' 간 피어슨 상관 분석

텍스트 분석항목	상관계수 r	유의확률 p-value
단락 당 문장 수	**-0.190788**	**5.974e-05**
텍스트 당 형태소 수	-0.415369	< 2.2e-16
단락 당 형태소 수	-0.510302	< 2.2e-16
문장 당 형태소 수	-0.624805	< 2.2e-16
한자비율(%)	-0.757976	< 2.2e-16
가나대 한자비율	-0.64965	< 2.2e-16
N5 출제기준 한자(%)	0.6808269	< 2.2e-16
N4 출제기준 한자(%)	-0.360835	6.855e-15
N3 출제기준 한자(%)	**-0.743337**	**< 2.2e-16**
N2 출제기준 한자(%)	-0.677393	< 2.2e-16
N1 출제기준 한자(%)	-0.555694	< 2.2e-16

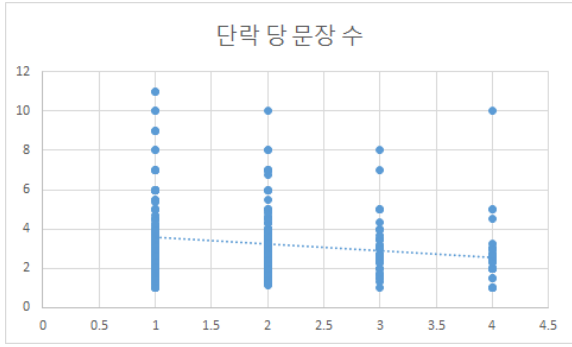

그림 2. '단락 당 문장 수'와 '가독성 레벨' 간의 상관관계를 나타내는 산포도와 추세선

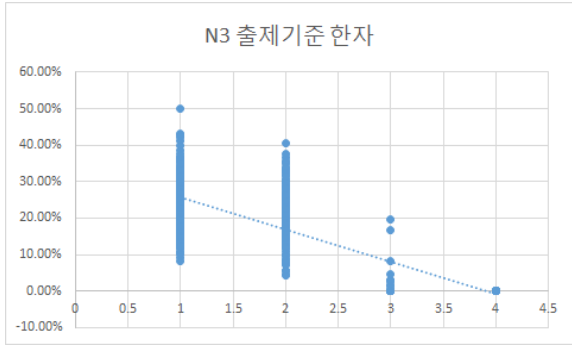

그림 3. 'N3 출제기준 한자(%)'와 '가독성 레벨' 간의 상관관계를 나타내는 산포도와 추세선

우선 표1에서 보는 바와 같이 텍스트 가독성에 영향을 미치는 요인일 것이라 가정한 각각

의 ‘텍스트 분석항목’과 ‘가독성 레벨’간의 상관관계가 상관분석의 결과를 통해 모두 통계적으로 유의미하다는 것을 확인할 수 있었다($p < .001$).

하지만 가독성 레벨과 텍스트 분석항목간의 상관관계가 통계적으로 유의미하다고 하더라도, 만일 각각의 상관계수 r의 값(e.g. 표1의 ‘한자비율(%)’의 상관계수 r = -.757976)이 크지 않다면 상관분석 결과가 현실적으로 활용할 수 있을 만큼 중요하지 않다고 말할 수 있다. 일반적으로 상관관계는 상관계수(Correlation Coefficient, r)를 사용하여 계량화 할 수 있는데, 모든 상관계수는 －1에서 +1 사이의 값이 된다. 상관계수의 절대 값의 크기는 상관성의 정도를 나타내며, 만일 상관계수가 0이라면 이는 두 변수 사이에 상관성이 없다는 것을 의미한다. 데이터의 성격에 따라 해석에 차이가 있을 수 있으나, 일반적으로 상관성이 얼마나 높은지를 양의 상관성을 기준으로 해석하면 다음의 (5)와 같다.

(5) [상관계수 판별 기준]
- a. $0 \leq r \leq 0.2$: 매우 낮은 상관성의 경우
- b. $0.2 < r \leq 0.4$: 낮은 상관성의 경우
- c. $0.4 < r \leq 0.7$: 비교적 높은 상관성의 경우
- d. $0.7 < r \leq 0.9$: 높은 상관성의 경우
- e. $0.9 < r \leq 1.0$: 매우 높은 상관성이거나 완전한 상관성의 경우

따라서 표1의 텍스트 분석항목 중, 상관계수 r의 값이 가장 큰 ‘N3 출제기준 한자(%)’의 경우, 그림2에서 보는 바와 같이 ‘N3 출제기준 한자’가 감소할수록 ‘가독성 레벨’ 또한 계속적으로 낮아진다. 이는 ‘N3 출제기준 한자’와 ‘가독성 레벨’ 사이에 상응하는 상관관계(r = -.743337, p < 2.2e-16)가 있다는 사실에 의하여 뒷받침 된다. 또한 상관계수 r값을 통해 두 변수가 ‘높은 상관성’을 가진다고 해석할 수 있다((5) 상관계수 판별 기준 참조). 그리고 그 외에도 텍스트 분석항목 중 ‘텍스트 당 형태소’는 ‘비교적 높은 상관성을 갖고 있다’고 볼 수 있다(r = -.415369, p < 2.2e-16). 하지만 ‘단락 당 문장 수’는 r = －.190788(p-value: 5.974e-05)이기 때문에, ‘단락 당 문장 수’와 ‘가독성 레벨 사이의 ‘상관성은 매우 낮다’고 볼 수 있다.

위와 같은 내용을 종합하면, 각각의 상관계수에 따라 그 정도의 차이는 있지만, 표1에서 제시된 ‘텍스트 분석항목’과 ‘가독성 레벨’ 사이에는 상응하는 상관관계가 있다고 할 수 있으며 이는 통계적으로 유의미하다.

4.2. 단순 선형회귀분석

이어서 본 절에서는 앞선 절에서 도출한 ‘텍스트 분석항목’이 ‘가독성 레벨’과 통계적으로 유의미한 상관성을 갖는다는 결론을 기반으로, 두 변수를 각각 독립변수X와 종속변수Y로 설정하여 독립변수의 변화에 따라 종속변수가 얼마나 영향을 받는지를 통계적으로 분석해보고자 한다. 이를 위해 다음의 두 가지 패턴의 ‘선형 회귀분석’ 모형을 설정하여 생각해 볼 수 있다.

(6) [선형 회귀 모형의 변수]

a. 독립변수 : 텍스트 분석 항목　　종속변수 : 가독성 레벨

b. 독립변수 : 가독성 레벨　　종속변수 : 텍스트 분석 항목

- 종속변수 : 독립변수 값의 변화에 따라 영향을 받는 변수

본고의 목적은 일본어 능력시험의 기출문제 분석을 통해 산출한 '가독성 레벨 공식'에 특정 일본어 '텍스트 분석항목'의 수치를 대입하여 해당 텍스트의 '가독성 레벨'을 예측하고자 하는 데에 있다. 즉, '텍스트 분석항목' 수치의 변화를 통해 가독성 레벨을 예측하고자 하므로 본고에서는 (6)a의 모형을 기본으로 선형 회귀분석을 실시하고자 한다. 우선 표1의 각각의 '텍스트 분석항목'을 독립변수, '가독성 레벨'을 종속변수로 하여 '단순 선형 회귀분석'을 실시했으며 그 결과는 이어지는 표2와 같다.

표 2. 가독성 레벨(종속변수 *Y*)과 텍스트 분석항목(독립변수 *X*) 간의 '단순 선형 회귀분석' (n=437)

텍스트 분석항목	회귀분석		분산분석	
	결정계수 R^2	조정 결정계수 AR^2	F-비	유의한 F
단락 당 문장 수	**0.0364002**	0.034185	16.43220305	5.974E-05
텍스트 당 형태소	**0.1725315**	0.170629	90.69976719	1.176E-19
단락 당 형태소	**0.2604079**	0.258708	153.1620603	2.386E-30
문장 당 형태소	**0.3903815**	0.38898	278.5610044	1.085E-48
한자비율(%)	**0.5745278**	0.57355	587.3935542	9.584E-83
가나대 한자비율	**0.4220446**	0.420716	317.6532304	9.559E-54
N5 출제기준 한자(%)	**0.4635253**	0.462292	375.8490134	8.418E-61
N4 출제기준 한자(%)	**0.1302019**	0.128202	65.11608915	6.954E-15
N3 출제기준 한자(%)	**0.5525499**	0.551521	537.1754707	5.59E-78
N2 출제기준 한자(%)	**0.4588617**	0.457618	368.8609832	5.559E-60
N1 출제기준 한자(%)	**0.3087958**	0.307207	194.3364971	8.908E-37

텍스트 분석항목	회귀분석 통계량				
	Y절편	계수	표준오차	t 통계량	p-value
단락 당 문장 수	2.2950862	-0.106654	0.0263106	-4.053665	**5.97371E-05**
텍스트 당 형태소	2.4442462	-0.002003	0.0002104	-9.523643	**1.17575E-19**
단락 당 형태소	2.68616	-0.0098	0.0007918	-12.37587	**2.38591E-30**
문장 당 형태소	3.3262508	-0.059391	0.0035585	-16.69015	**1.0847E-48**
한자비율(%)	3.7298028	-7.902216	0.3260501	-24.2362	**9.58371E-83**
가나대 한자비율	3.1046145	-3.17701	0.1782551	-17.82283	**9.55944E-54**
N5 출제기준 한자(%)	0.8179701	2.784464	0.1436266	19.386826	**8.41844E-61**
N4 출제기준 한자(%)	2.6557945	-2.796368	0.3465374	-8.069454	**6.95412E-15**
N3 출제기준 한자(%)	3.0412783	-6.292523	0.2714981	-23.17705	**5.58996E-78**
N2 출제기준 한자(%)	2.8190747	-8.618219	0.4487311	-19.20575	**5.55901E-60**
N1 출제기준 한자(%)	2.5209287	-13.19557	0.9465659	-13.94046	**8.90824E-37**

표2의 결과를 통해, 모든 '텍스트 분석항목'의 Y절편과 계수의 t-통계량에 대한 p-value가

0.001보다도 작기 때문에($p < .001$) 각각의 Y절편 및 계수가 통계적으로 유의미하다는 것을 확인할 수 있다. 또한, 분산분석의 'F-비' 값 또한 '유의한 F 값'을 상회하고 있기 때문에 Y절편과 계수로 구성된 각각의 단순 선형 회귀모형 또한 통계적으로 유의미하다.

한편 '단순 선형 회귀분석'의 결정계수 $\mathbf{R^2}$는 회귀식으로 설명할 수 있는 회귀모형의 적합도를 나타내는데, 0과 1사이의 값을 가지며 1에 가까울수록 해당 회귀모형의 적합도가 높다는 것을 의미한다. 예를 들어, 표2의 '텍스트 분석항목' 중 '한자비율(%)'의 $\mathbf{R^2}$ = .5745278을 통해 '한자비율(%)'을 독립변수로 하고 '가독성 레벨'을 종속변수로 하는 회귀모형이 약 57%의 비율로 적합도를 갖는 다는 것을 알 수 있다. $\mathbf{R^2}$의 값에는 절대적 척도가 있는 것은 아니기 때문에, 본고에서는 다음의 (7)과 같은 Cohen(1988)과 홍정하 외(2011)의 판별기준에 따라 각각의 $\mathbf{R^2}$값을 판단하고자 한다.

(7) [$\mathbf{R^2}$값의 판별]
  a. 0.01 : 낮은 효과
  b. 0.09 : 중간 효과
  c. 0.25 : 높은 효과

표2의 결과를 (7)의 기준을 통해 보자면 '단락 당 문장 수'의 경우 '낮은 효과'로, '텍스트 당 형태소'와 'N4 출제기준 한자(%)'가 '중간 효과'로, 그 외의 '텍스트 분석항목'은 모두 '높은 효과'로 판별할 수 있다. 다시 말하면, 단락 당 문장 수'와 '텍스트 당 형태소', 'N4 출제기준 한자(%)'의 경우 결정계수R^2 값이 작기 때문에 통계적 예측력이 상대적으로 약하다고 할 수 있지만, 그 외의 텍스트 분석 항목의 통계적 예측력은 상대적으로 강하다고 할 수 있다.

4.3. 다중 선형회귀 분석

지금까지 '상관관계 분석'과 '단순 선형회귀 분석'을 통해 가독성 레벨에 영향을 미치는 텍스트 분석항목의 통계적 유의성을 검토해 보았다. 이를 통해, 다소간의 편차가 있으나 본고에서 설정한 각각의 텍스트 분석항목은 모두 통계적 유의성을 보이고 있다는 점을 확인할 수 있었다. 본 절에서는 이와 같은 개별적 일본어 텍스트 분석항목과 가독성 레벨과의 상관성 분석을 바탕으로, 이들 유의성을 보이는 복수의 '텍스트 분석항목'으로 구성된 선형회귀모형을 가지고 분석을 실시하고자 한다. 이와 같은 선형회귀 모형 분석을 '다중 선형회귀 분석'이라고 하는데, 이를 위해서는 우선 복수의 '텍스트 분석항목'을 합하여 '독립변수'로 설정하고 이를 통해 영향을 받는 '종속변수'로 '가독성 레벨'을 설정한 뒤, 가독성 레벨 분석을 위한 최적의 선형회귀 모형을 설정할 필요가 있다.

본고에서는 R 통계 패키지를 통해 AIC(Akaike information criterion, Akaike(1974))를 활용하여 표1과 표2에서 제시한 텍스트 분석항목에서 제시된 변수를 사용하여 표3과 같은 결과를 도출했으며 이를 통해 이어지는 (8)과 같은 다중 선형회귀 모형을 얻을 수 있었다.

표 3. ‘R 통계 패키지’를 통한 AIC 결과화면 중 일부 – step( ) 함수

Start: **AIC=-713.41**
가독성레벨 ~ 단락당문장 + 텍스트당형태소 + 단락당형태소 + 문장당형태소 + 한자. + 가나대한자비율 + N5한자. + N4한자. + N3한자. + N2한자. + N1한자.

	Df	Sum of Sq	RSS	AIC
- 단락당형태소	1	0.0628	80.904	-715.07
- N5한자.	1	0.0806	80.922	-714.98
- 단락당문장	1	0.2683	81.110	-713.96
<none>			80.841	-713.41
- 텍스트당형태소	1	1.3895	82.231	-707.97
- 문장당형태소	1	1.7303	82.572	-706.16
- 가나대한자비율	1	2.8254	83.667	-700.40
- N4한자.	1	4.9382	85.780	-689.50
- N2한자.	1	5.3780	86.219	-687.27
- 한자.	1	5.7563	86.598	-685.35
- N3한자.	1	6.5362	87.378	-681.44
- N1한자.	1	7.1990	88.041	-678.13

Step: **AIC=-715.07**
가독성레벨 ~ 단락당문장 + 텍스트당형태소 + 문장당형태소 + 한자. + 가나대한자비율 + N5한자. + N4한자. + N3한자. + N2한자. + N1한자.

…… 중략 ……

Call:
lm(formula = **텍스트레벨 ~ 단락당문장 + 텍스트당형태소 + 문장당형태소 + 한자. + 가나대한자비율 + N4한자. + N3한자. + N2한자. + N1한자.,**
data = p1_all)

Residuals:

Min	1Q	Median	3Q	Max
-1.43179	-0.28678	0.04499	0.31363	0.98111

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.2595618	0.0822339	51.798	< 2e-16	***
단락당문장	-0.0234703	0.0127543	-1.840	**0.066435**	.
텍스트당형태소	-0.0003099	0.0001161	-2.668	**0.007910**	**
문장당형태소	-0.0125436	0.0027826	-4.508	8.47e-06	***
한자.	-0.0609040	0.0109462	-5.564	4.66e-08	***
가나대한자비율	1.6948966	0.4439320	3.818	0.000155	***
N4한자.	-0.0129798	0.0018617	-6.972	1.19e-11	***
N3한자.	-0.0209141	0.0027868	-7.505	3.61e-13	***
N2한자.	-0.0241039	0.0038283	-6.296	7.58e-10	***
N1한자.	-0.0402633	0.0060682	-6.635	9.85e-11	***

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

Residual standard error: 0.4355 on 427 degrees of freedom
Multiple R-squared: 0.8014, **Adjusted R-squared: 0.7972**
F-statistic: 191.5 on 9 and 427 DF, **p-value: < 2.2e-16**

(8) AIC기반 다중 선형회귀 모형 설정

a. 입력변수

- 종속변수 : 가독성 레벨 (1～4)
- 독립변수 : 단락 당 문장 수, 텍스트 당 형태소, 단락 당 형태소, 문장 당 형태소, 한자비율(%), 가나대 한자비율, N5 출제기준 한자(%), N4 출제기준 한자(%), N3 출제기준 한자(%), N2 출제기준 한자(%), N1 출제기준 한자(%)

b. AIC 분석 결과 - 다중 선형회귀 모형

- [단락 당 문장 수 + 텍스트 당 형태소 + 문장 당 형태소 + 한자비율(%) + 가나대 한자비율 + N4 출제기준 한자(%) + N3 출제기준 한자(%) + N2 출제기준 한자(%) + N1 출제기준 한자(%)]

위에서 보는 바와 같이 AIC를 활용하여 선형회귀 모형을 설정한 결과, 텍스트 분석항목 중 '단락 당 형태소'와 'N5 출제기준 한자(%)'를 다중 회귀분석의 독립변수에서 제외할 수 있었는데, 해당 모형은 조정 결정계수 AR^2 = .7972로, 독립변수들이 종속변수 '텍스트 레벨'의 변동을 약 80% 정도 설명할 수 있으며 이는 매우 높은 효과라고 할 수 있다((7) 참조). 그리고 분산분석 결과 p-value가 0.001보다 작아(p-value: < 2.2e-16, F-statistic: 191.5), 본 회귀식이 통계적으로 유의미하다는 것을 알 수 있다.

위와 같은 분석을 통해 다음의 (9)와 같은 회귀 공식을 얻을 수 있는데, 이것이 '가독성 레벨'을 산출하는 공식이 된다.

(9) [일본어 텍스트의 가독성 레벨 공식]

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$$

$$y = 4.2595618 - 0.0234703 x_1 - 0.0003099 x_2 - 0.0125436 x_3 - 0.0609040 x_4 + 1.6948966 x_5 - 0.0129798 x_6 - 0.0209141 x_7 - 0.0241039 x_8 - 0.0402633 x_9$$

- y: 일본어 텍스트 레벨
- α: Y절편, $\beta_1 \sim \beta_9$: 각 독립변수의 회귀계수
- x_1: 단락 당 문장 수, x_2: 텍스트 당 형태소, x_3: 문장 당 형태소, x_4:한자비율, x_5:가나대 한자비율, x_6:N4급 출제기준 한자, x_7:N3 출제기준 한자, x_8:N2 출제기준 한자, x_9:N1 출제기준 한자

그런데 표3과 (9)의 x_1의 회귀계수, 즉 '단락 당 문장 수'의 회귀계수(-0.0234703)는 p-value가 유의수준 0.05에서 통계적으로 유의미하지 않은 결과(p-value: .066435)를 보이고 있다. 이는 해당 모형에서 '단락 당 문장 수'가 '가독성 레벨' 변화의 요인으로 보기 어렵다는 것을 나타낸다. 그리고 '텍스트 당 형태소'의 경우 유의수준 0.001에서 유의미한 결과(p-value: .007910)를 보이고는 있으나 타 독립변수와 다소 차이를 보이고 있으며 회귀계수 또한 낮다(-0.0003099). 이에 본고에서는 (8)의 선형회귀 모형에서 '단락 당 문장 수'와 '텍스트 당 형태소'를 제외한 다중 선형회귀 분석을 통해, 표4와 같은 결과를 얻을 수 있었다.

표 4. ‘R 통계 패키지’를 통한 다중 선형회귀 분석

```
Call:
lm(formula = 가독성레벨 ~ ., data = p1_aic_del12)
Residuals:
      Min        1Q    Median        3Q       Max
-1.32692 -0.29497  0.03326  0.30555  1.04934
Coefficients:
                Estimate   Std. Error  t value Pr(>|t|)
(Intercept)     4.188123   0.075155    55.726  < 2e-16 ***
문장당형태소    -0.011641  0.002755    -4.225  2.92e-05 ***
한자.           -0.066788  0.010885    -6.136  1.93e-09 ***
가나대한자비율   1.927587  0.441795     4.363  1.61e-05 ***
N4한자.         -0.013479  0.001871    -7.203  2.67e-12 ***
N3한자.         -0.022488  0.002745    -8.193  2.97e-15 ***
N2한자.         -0.024522  0.003861    -6.351  5.46e-10 ***
N1한자.         -0.042728  0.005989    -7.134  4.17e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4395 on 429 degrees of freedom
Multiple R-squared:  0.7968,    Adjusted R-squared:  0.7935
F-statistic: 240.3 on 7 and 429 DF,  p-value: < 2.2e-16
```

표4의 선형회귀 모형은 조정 결정계수 AR^2 = .7935로, 독립변수들이 종속변수 ‘텍스트 레벨’의 변동을 약 80% 정도 설명할 수 있다. 그리고 분산분석 결과 p-value가 0.001보다 작아 (p-value: < 2.2e-16, F-statistic: 240.3), 본 회귀식이 통계적으로도 유의미하다. 또한 모든 독립변수의 p-value가 모두 0.001보다 작기 때문에 모든 독립변수의 회귀계수가 통계적으로 매우 유의미하기 때문에 잠정적으로 다음과 같은 ‘다중 선형회귀 모형 I’을 설정할 수 있다.

(10) [다중 선형회귀 모형 I]

a. 종속변수 : 가독성 레벨 (1~4)

b. 독립변수

- [문장 당 형태소 + 한자비율(%) + 가나대 한자비율 + N4 출제기준 한자(%) + N3 출제기준 한자(%) + N2 출제기준 한자(%) + N1 출제기준 한자(%)]

4.3.1 다중 공선성 진단

앞선 절에서 실시한 다중 선형회귀 분석이 유효하기 위해서는 전제가 필요한데, 각 독립변수들 간에 선형종속관계 혹은 유사한 관계가 성립해서는 안 된다는 점이다. 이와 같은 현상을 다중 공선성의 문제라고 하는데, 다중 공선성이 심할 경우, 회귀분석의 결과를 왜곡시켜 특정 독립변수의 종속변수에 대한 효과를 측정하는 데에 문제를 야기한다. 보통 이와 같은 다중 공선성 문제가 있는지 여부를 진단하는 지표로는 공차Tolerance와 VIF(Variance Inflation Factor)를 들 수 있는데, 본고에서는 VIF지표를 통해 위 (10)의 '다중 선형회귀 모형 I'의 각 독립변수들 간에 다중 공선성 문제가 존재하는지 여부를 확인했으며, 그 결과는 표5와 같다.

표 5. ‘R 통계 패키지’를 통한 회귀모형 I (10)에 대한 다중 공선성 진단 – vif( )

독립변수	문장 당 형태소	한자비율(%)	가나대 한자비율	
VIF	1.773901	**23.015857**	**17.230884**	
독립변수	N4출제기준한자(%)	N3출제기준한자(%)	N2출제기준한자(%)	N1출제기준한자(%)
VIF	1.231070	2.219760	1.944533	1.342970

VIF 값이 10이상 이면 독립변수들 간에 다중공선성이 존재한다고 판단하는데, 위 표5의 결과를 통해, 독립변수 중 ‘한자비율(%)’과 ‘가나대 한자비율’의 독립변수에 다중 공선성이 존재한다는 것을 알 수 있다. 일반적으로 다중 공선성이 발생하면 이를 해결하기 위해서 1) 공선성을 발생시킨 변수를 제거하거나 2)표본의 크기를 키우거나, 3)변수를 변환시키거나, 4) 편의추정법을 사용하는 등의 방법을 사용한다. 본고에서는 다중 선형회귀 모형 I (10)에서 독립변수 ‘한자비율(%)’와 ‘가나대 한자비율’을 하나씩 제거해 보는 것을 통해 통계적으로도 유의미하면서도 AR^2의 값이 상대적으로 큰 모형을 설정하고, 다시금 다중 공선성 진단을 실시했는데, 그 과정 및 결과는 다음과 같다.

(11) [다중 공선성 진단 결과 반영 – 다중 선형회귀 모형 II]

a. 종속변수 : 가독성 레벨

b. (10)b의 독립변수 중 ‘한자비율(%)’ 제거

- [텍스트 당 형태소 + 문장 당 형태소 + 가나대 한자비율 + N4 출제기준 한자(%) + N3 출제기준 한자(%) + N2 출제기준 한자(%) + N1 출제기준 한자(%)]
- Residual standard error: 0.4578 on 430 degrees of freedom
  Multiple R-squared: 0.779, **Adjusted R-squared: 0.7759**
  F-statistic: 252.6 on 6 and 430 DF, **p-value: < 2.2e-16**

c. (10)b의 독립변수 중 ‘가나대 한자비율’ 제거

- [텍스트 당 형태소 + 문장 당 형태소 + 한자비율(%) + N4 출제기준 한자(%) + N3 출제기준 한자(%) + N2 출제기준 한자(%) + N1 출제기준 한자(%)]
- Residual standard error: 0.4486 on 430 degrees of freedom
  Multiple R-squared: 0.7878, **Adjusted R-squared: 0.7848**
  F-statistic: 266 on 6 and 430 DF, **p-value: < 2.2e-16**
- 회기계수 통계량: Estimate Std. Error t value Pr(>|t|)

(Intercept)	**4.041029**	0.068564	58.938	< 2e-16 ***
문장당형태소	-0.011292	0.002811	-4.017	6.97e-05 ***
한자.	-0.022071	0.003743	-5.896	7.54e-09 ***
N4한자.	-0.016339	0.001789	-9.132	< 2e-16 ***
N3한자.	-0.025853	0.002689	-9.615	< 2e-16 ***
N2한자.	-0.026349	0.003918	-6.725	5.60e-11 ***
N1한자.	-0.046882	0.006036	-7.768	5.92e-14 ***

표 6. ‘R 통계 패키지’를 통한 회귀모형 II (11)c에 대한 다중 공선성 진단 - vif( )

독립변수	문장 당 형태소	가나대 한자비율	N4출제기준한자(%)
VIF	1.763450	1.955818	1.038984
독립변수	N3출제기준한자(%)	N2출제기준한자(%)	N1출제기준한자(%)
VIF	1.918614	1.847129	1.294415

앞선 (10)b에서 독립변수 중 ‘가나대 한자비율’을 제거한 (11)c의 다중 회귀진단 모형은 (11)b와 비교하여 AR^2의 값이 상대적으로 크고, 다중 공선성 진단 결과도 위 표6과 같이 모든 독립변수의 VIF 값이 10미만이기 때문에 다중 공선성 문제가 발생하지 않는다. 또한 각 독립변수의 p-value가 모두 0.001보다 작아, 모든 독립변수의 회귀계수가 통계적으로 매우 유의하다. 이와 같은 결과를 반영하여 다음과 같은 일본어 텍스트의 가독성 공식을 산출했다.

(12) [일본어 텍스트의 가독성 레벨 공식] - AR^2: .7848, p-value: < 2.2e-16

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

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

- y:일본어 텍스트 레벨
- α: Y절편, $\beta_1 \sim \beta_9$: 각 독립변수의 회귀계수
- x_1: 문장 당 형태소, x_2: 한자비율(%), x_3: N4급 출제기준 한자, x_4: N3 출제기준 한자, x_5: N2 출제기준 한자, x_6: N1 출제기준 한자

4.3. 일본어 텍스트 가독성 공식 검증(회귀 진단Regression diagnostics)

전 절에서는 (12)와 같이 다중 회귀분석을 통해 일본어 텍스트의 ‘가독성 레벨 공식’을 산출할 수 있었다. 그런데 회귀분석은 독립변수를 가지고 종속변수의 변화를 예측하고자 하는 것으로, 종속변수의 실제값과 회귀식에 의한 기댓값은 차이가 나기 마련인데 통계학에서는 이 차이를 잔차Residual라고 한다. 따라서 어디까지나 가정인 ‘가독성 레벨 공식’을 받아들이기 위해서는 다음과 같은 잔차에 대한 가정이 성립해야만 한다.

(13) 잔차의 가정

a. 선형성 : 잔차의 평균 = 0
b. 등분산성 : 잔차는 등분산이어야 한다.
c. 독립성 : 잔차는 서로 독립적이다.
d. 정규성 : 잔차는 정규분포를 따른다.

우선 (12)의 회귀식의 잔차에 대한 (13)의 항목을 차례대로 분석하기 위해 잔차도를 그려보면 다음의 그림4와 같다.

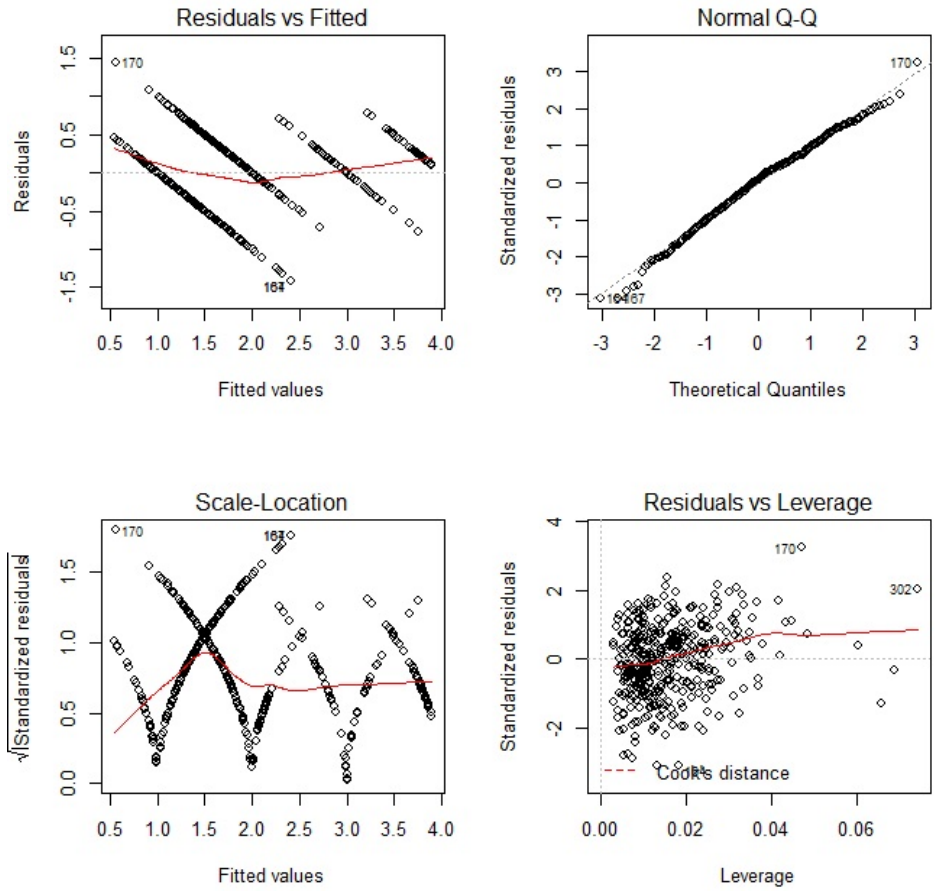

그림 4. (12)일본어 텍스트의 가독성 레벨 공식의 잔차도

그림4를 차례로 분석해 보자면 우선, 좌측 상단의 '잔차vs적합값(Residuals vs Fitted)' 그래프를 보면 잔차가 0 주변에 흩어져 있어 (12)의 가독성 공식(회귀공식)은 (13)a선형성을 만족하는 것으로 보이지만 약간의 패턴이 보이기 때문에 보다 상세한 검증이 필요할 것으로 보인다. 한편, 좌측 하단의 '표준화 잔차의 제곱근과 적합값(standrdized residual and Fitted values)'이 직선에 가까우므로 (13)b등분산성 가정에는 문제가 없어 보이며, 우측 상단의 '정규확률 그래프(Normal Q-Q Plot)'가 직선에 가까우므로 (13)d정규성 또한 만족할 것으로 보인다. 그리고 마지막으로 우측 하단의 '잔차vs레버러지(Residuals vs Leverage)' 그래프를 통해 이상치outlier가 될 가능성이 있는 값으로 170과 302번째 데이터가 있을 수 있다는 점을 확인할 수 있었다.

4.3.1 잔차의 선형성 진단과 등분산성 테스트

본 절에서는 상기 그림4의 대략적인 잔차도 분석을 기반으로 더욱 상세한 검증을 차례로 실시해 보고자 한다. 우선 잔차의 평균값이 -4.68534E-17로 0에 매우 가깝기 때문에 그림4와 종합해 판단할 때, (12)의 가독성 공식은 (13)a의 선형성을 만족한다고 할 수 있다.

이어서 등분산성 검증을 위해, 브로슈-파간 테스트(Breusch- Pagan Test)을 실시했으며 그 결과는 다음의 표7과 같다. 즉, p-value = .1776으로 유의수준 0.05에서 귀무가설 지지하기 때문에 등분산이라는 것을 알 수 있다.

표 7. 'R 통계 패키지'를 통한 가독성 레벨 공식 (12)에 대한 브로슈-파간 테스트 – bptest( ) 함수

studentized Breusch-Pagan test data: result_aic_del12_c BP = 8.9291, df = 6, **p-value = 0.1776**

4.3.2 잔차의 독립성 진단과 정규성 진단

이어서 잔차의 독립성과 정규성을 진단하기 위해서 더빈-왓슨 테스트(Durbin-Watson test)와 샤피로-윌크 정규성 테스트(Shapiro-Wilk normality test)를 실시했으며, 그 결과는 다음의 표8·9과 같다.

표 8. 'R 통계 패키지'를 통한 가독성 레벨 공식 (12)에 대한 더빈-왓슨 테스트 – dwtest( )

data: result_aic_del12_c **DW = 1.1327**, p-value < 2.2e-16 alternative hypothesis: true autocorrelation is greater than 0

표 9. 'R 통계 패키지'를 통한 가독성 레벨 공식 (12)에 대한 샤피로-윌크 테스트 – shapiro.test( )

data: resid(result_aic_del12_c) W = 0.9922, **p-value = 0.0218**

더빈-왓슨 값은 0과 4 사이에 존재하는데, 0의 경우 양의 자기상관을, 4의 경우 음의 자기상관을 나타낸다[4]. (12)의 가독성 공식은 표9와 같이 dw = 1.1327로 0보다 2에 가깝기 때문에 대략적인 (13)c독립성을 판단할 수 있다.

한편, 그림4의 우측 상단의 그래프를 통해서 정규성을 만족할 것으로 보았으나, 표9의 결과를 보면, p-value = .0218로 유의확률 0.05에서 귀무가설이 기각되어 정규성을 만족할 수 없다. 하지만, 유의확률 0.01에서는 정규성을 만족할 수 있다. 그리고 회귀분석에서 잔차의 정규분포 여부는 등분산과 같이 엄격하게 적용되지 않는다. 또한 모집단의 정규분포가 정규분포 한다는 가정이 성립되면, 그 표본의 분포가 정규분포에 약간 위배되더라도 회귀분석에 크게 문제를 일으키지 않으며, 표본의 크기가 크면(일반적으로 30 이상) 중심극한정리에 의해 정규분포이론을 적용할 수 있다고 알려져 있다.

4) 구체적으로는 더빈왓슨 표에서 dw 통계량의 하한치(dL)과 상한치(dU)를 투입된 독립변수의 수(k)와 표본의 사례수(n)를 통해 구한 후 dw 통계량이 dU보다 크고 4-dU보다 작으면 자기상관 없다고 판단한다.

4.3.2. 이상치 검정

이상치Outlier란 회귀모형에 의해 설명되지 않는 특이한 데이터로, 이를 검출하기 위해서는 스튜던트화 잔차(studentized residual)를 사용한다. 그림4의 우측 하단의 '잔차vs레버러지(Residuals vs Leverage)' 그래프를 통해 확인한 바와 같이 170과 302번째 데이터가 이상치가 될 가능성이 있기 때문에, 이상치 검정을 실시했으며 그 결과는 다음의 표10과 같다. 그러나 Bonferonni = .50302로 수치가 유의수준 0.05보다 크므로 이상치가 하나도 없다는 결론을 내릴 수 있었다.

표 10. 'R 통계 패키지'를 통한 가독성 레벨 공식 (12)에 대한 이상치 검정 – outlierTest( )

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
rstudent unadjusted p-value **Bonferonni p**
170 3.272783 0.0011511 **0.50302**

이상과 같이 '가독성 레벨 공식'에 대한 회귀 모형 진단을 통해 (12)일본어 텍스트 가독성 레벨 공식을 통계적으로 유의하다고 판단하고 본고의 최종 공식으로 확정할 수 있었다.

5. 맺음말

위와 같이 구JLPT 독해지문을 바탕으로 가독성과 상관관계가 있다고 판단되는 텍스트 분석항목을 도출하고 이를 통해 다음의 (14)와 같이 유의한 독립변수를 설정하고 이를 사용하여 가독성 레벨 공식을 도출해 낼 수 있었다. 참고로 (14)의 가독성 공식은 일본어 가독성 분석 툴인 「AJ-JpnRa Tool[5)]」에 반영되어 배포 중에 있으므로 해당 프로그램을 통해 언제든지 손쉽게 일본어 텍스트의 가독성 분석이 가능하다.

(14) [일본어 텍스트의 가독성 레벨 공식] - AR^2: 0.7848, p-value: < 2.2e-16

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

- y: 일본어 텍스트 레벨
- x_1: 문장 당 형태소, x_2: 한자비율(%), x_3: N4급 출제기준 한자, x_4: N3 출제기준 한자, x_5: N2 출제기준 한자, x_6: N1 출제기준 한자

5) 공개용 일본어 텍스트 가독성 분석 프로그램으로, 자세한 프로그램 설명 및 사용법은 김유영(2014) 및 다음의 웹페이지를 참조할 것.
- http://japanese.or.kr/japaneseutill/Readability%20Tool%20Series/About%20AJ-Readability_Tool.aspx

물론, Davison(1986:28)이 언급한 바와 같이 '텍스트 분석항목'과 '가독성 레벨' 간의 관계는 인과관계가 아니라 어디까지나 상관관계이기 때문에, 본고의 가독성 분석 공식이 가독성 레벨이 높아지는 원인을 설명하고 있는 것은 아니다. 그리고 텍스트의 가독성에는 '리더 Reader 영역'과 같이 인간의 직관으로 판단할 수 있는 영역이 있는 것도 사실이다. 따라서 저자는 금번 연구 결과를 바탕으로 다음 연구에서는 가독성에 영향을 미치는 텍스트 영역 중 어휘 및 문법의 난이도 등에 대한 분석을 추가하여 더욱 정교한 가독성 레벨 판단 공식을 산출하고자 한다. 이를 통해 개발된 일본어 텍스트의 가독성 공식은 다양한 영역에서 일본어 텍스트의 가독성을 판별하는 하나의 지표가 될 수 있을 것으로 전망한다.

◀ 참고문헌(Reference) ▶

김유영(2013)「동일본대지진과 일본사회의 언어—외국인 주민을 대상으로 하는 '공적 정보제공'에 있어서의 '알기 쉬운 일본어やさしい日本語'—」『일본학보』97집, 한국일본학회, Yu Young, Kim(2013)「2011 Tōhoku earthquake and tsunami and Language of Japan; 'Easy Japanese' in 'Public information offering' for 'Foreign-born population'」『The Korean Journal of Japanology』Vol.97, pp.17-38.

김유영(2014)「일본어 텍스트 가독성 분석—舊JLPT 기출문제 기반 일본어텍스트의 한자 및 한자 어휘 가독성 판단 프로그램의 개발—」『日本言語文化』27호 한국일본언어문화학회, Yu Young, Kim(2014)「An analytical Study on the Readability of Japanese text—With a focus on Developing an Japanese text analysis Program Based on Previous format of JLPT—」『Journal of Japanese Language & Culture』Vol.27, pp.357-382.

홍정하 외(2011)「텍스트 수준과 가독성 : 한국어 학습 교재를 이용한 검증과 응용」『언어정보』Vol.12, 고려대학교 언어정보연구소, Jeong Ha, Hong and Jae Woong, Choi and Seok Hoon, Yoo(2011)「A Verification and Application of a Correlation between Text Levels and Readability Using Korean Learning Materials」『Language Information』Vol.12, pp.111-148.

황영희・김유영(2014)「온라인대학 일본어 기초과정의 한자 콘텐츠-일본어능력시험 기출한자 및 상용한자와의 대조-」『일본어학연구』 40호, 한국일본어학회, Young Hee, Hwang and Yu Young, Kim(2014)「Education Contents for Japanese Chinese characters of Online University‐Contrastive research of regular-use Chinese characters and JLPT」『Journal of Japanese Language』Vol.40, pp.165-180.

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, Tateisi Yuka, ONO Yoshihiko, Yamada Hisao(1988)「Derivation of a Readability Formula of Japanese Texts」『IPSJ SIG technical reports』1988-HI-018, 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, INUI Hiroko and OKADA Naoyuki(2000)「Is a Long Sentence Always Incomprehensible? : A Structural Analysis of Readability Factors」『IPSJ SIG technical reports』2000(11), pp.63-70.

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, TAKAGI Yuko(1991)「Toward Development of Rapid Reading Material-on the Basis of the Research of Readability」『Papers in teaching Japanese as a foreign language』Vol.1, pp.66-85.

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, MANTO Katsunori(2000)「Appraisal of Readability Formulas in Japan's ELT」『English review』Vol.15, 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, Motooka Naoko(1987)「A Study on Readability : From the Viewpoint of Schema Theory(Papers Read at Shimane Convention)」『CASELE research bulletin』17, pp.1-8.

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, Kitao Kenji and Kitao S. Kathleen(2011)『Journal of culture and information science』6(1), pp.11-20.

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, Yamada Jun and Vernon Charles(1977) 「An Investigation of Readability (Papers read at Nagoya Convention)」『CASELE research bulletin』(7). pp.79-84.
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. KOJIMA Kensuke and SATO Satoshi(2009) 「Readability Assignment to Balanced Corpus of Contemporary Written Japanese」『IEICE technical report』 109(84), pp.13-18.
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, KOJIMA Kensuke and Satoshi Sato and FUJITA Atsushi(2009) 「Estimation of Japanese text Complexity based on Moji bigram model」『Journal of language processing』 15, 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.
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, SHIBASAKI Hideko and TAMAOKA Katsuo(2010)「Constructing a Formula to Predict School Grades 1-9 based on Japanese Language School Textbooks」『Japan journal of educational technology』 33(4), pp.449-458.
日本国語大辞典編集委員会編(2001)『日本国語大辞典』 第2版, 小学館, Edition staff of Nihon Kokugo Daijiten(2001) 『the Nihon Kokugo Daijiten』 2nd edition
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, Tanaka Kenji(1996)「A New Objective Readability Yardstick」『The Jacet ... Annual Convention』 35, pp.347-348.
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. So Kokui and KANO Toshiyuki(2010)「構文的な分かりやすさを評価する可読性評価技術」『Journal of language processing』 16, pp.51-55.
佐藤理史(2008)「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, SATO Satoshi(2008)「Estimation of Japanese text Complexity」『Language(Gekkan Gengo)』 37(8), pp.54-57.
佐藤理史(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). SATO Satoshi(2011) 「Measuring Text Readability Based on Balanced Corpus」『IPSJ Journal』, 52(4), pp.1777-1789.
佐藤理史(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, SATO Satoshi(2013) 「Text Readability and Word Distribution」『IPSJ SIG technical reports』 2013-NL-213(6), pp.1-11.
佐藤理史(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. SATO Satoshi(2014)「Measuring Japanese Language Proficiency of Computer by Using University Entrance Examination」『Japio year book』 pp.274-277.
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, KAWAMURA Yoshiko(1998) 「A computer system for checking the level of difficulty of Japanese reading material : Development and uses of the Vocabulary Level Checker」『Japanese language education methods』 5-2, 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(2013), Canadian Association for Japanese Language Education. KAWAMURA Yoshiko(2013) 「Implementation and Evaluation of a System for Determining Japanese Text-Level Difficulty」『Journal CAJLE』 pp.18-30.
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, Kiyokawa Hideo(1978) 「A Research Review on Readability Studies」『Shukutoku University bulletin』 12, pp.65-82.
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, TSUCHIYA Takehisa(1998) 「A Metrical Analysis of College-level Reading Materials and its Implications for Readability」『The Jacet ... Annual Convention』 37, pp.50-51.
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, HIGUCHI Akihiko and HIGUCHI Takako(2012) 『Bulletin of the Faculty of Education, Kagoshima University. Studies in education』 64, 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, Hiramoto Satoshi(1994)「A Study on Factors of Text Readability : A Case for Discourse Readability(Papers Read at Tottori Convention)」『CASELE research bulletin』(24), 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, Sangawa Kristina Hmeljak(2014)「Construction and analysis of a corpus of Japanese textbooks as a basis for the development of a readability measurement tool for learners of Japanese as a foreign language」『Journal of Japanese language education methods』 19(2), pp.4-5.

Akaike, H. 1974. A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19-6, pp.716-723.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), New York: Academic Press.

Dale, E. & Chall, J. S. 1948. "A formula for predicting readability." *Educational Research Bulletin* 27, 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.

Flesch, R. 1948. "A New Readability Yardstick." *Journal of Applied Psychology* 32, pp.221-233.

Gilliland, J. 1972, *Readability*, University of London Press.

Klare, G. R. 1963. *The measurement of readability*. T. Ames, IA:Iowa State University Press, pp.1-358.

Oxford University Press. 2010. *Oxford Advanced Learner's Dictionary*. Oxford University Press.

Spache, G. 1953. "A new readability formula for primary grade reading materials." *Elementary School Journal* 53-7, pp.410-413.

<要 旨>

텍스트 가독성은 텍스트의 수준 및 난이도를 평가하는 통계적 척도로, 다양한 요인에 의해 영향을 받는다. 본고에서는 이러한 일본어 텍스트의 가독성 레벨에 영향을 미치리라 생각되는 다양한 요인들 중에서도 텍스트 영역에서 수치화 할 수 있는 요인들만을 추출하여 이를 통계적 기법을 통해 분석했다. 이를 통해 각각의 요인들의 유효성을 검증하고, 이를 바탕으로 가독성 레벨 판정 시스템의 핵심이 되는 가독성 레벨 공식을 개발했다.

구체적으로 최근 20년간의 구舊JLPT(1급~4급) 독해・문법 영역 중 독해 지문만을 선별하여 구축한 'JLPT 독해 지문 데이터베이스'를 기반으로 피어스 상관분석을 수행하여, 텍스트의 '문장 당 형태소', '한자비율(%)', 'N4급 출제기준 한자 비율', 'N3 출제기준 한자 비율', 'N2 출제기준 한자 비율', 'N1 출제기준 한자 비율'을 변수로 설정하고, 다시금 다중 선형회귀 분석을 통해 다음과 같은 일본어 텍스트 가독성 공식을 산출했다.

[일본어 텍스트의 가독성 레벨 공식] - AR^2: .7848, p-value: < 2.2e-16

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

- y: 일본어 텍스트 레벨
- x_1: 문장 당 형태소 수, x_2: 한자비율(%), x_3: N4급 출제기준 한자비율(%),
  x_4: N3 출제기준 한자비율(%), x_5: N2 출제기준 한자비율(%), x_6: N1 출제기준 한자비율(%)

주제어 : 가독성, 가독성 공식, 텍스트 레벨, 일본어 텍스트 가독성지표, 코퍼스, 일본어능력시험, 상관분석, 선형회귀분석, 가독성 분석기, AJ-JpnRa Tool

■ 투 고 : 2015. 02. 28.
■ 심 사 : 2015. 03. 15.
■ 심사완료 : 2015. 04. 20.