텍스트 수준과 가독성:
한국어 학습 교재를 이용한 검증과 응용*

홍정하 · 최재웅 · 유석훈**

Jungha Hong, Jae-Woong Choe & Seok-Hoon You. 2011. **A Verification and Application of a Correlation between Text Levels and Readability Using Korean Learning Materials**. *Language Information*. 111-148. Readability is commonly defined as a statistical scale for measuring the ease of reading and understanding a text. In this paper, we propose that paragraph length, sentence length, and word length are applicable to readability metrics for Korean texts. A correlation between these readability metrics and text levels is statistically analyzed using correlation analysis and simple and multiple linear regression analyses based on data as found in Korean learning materials. Paragraph length and sentence length show a steady upward slope with increasingly higher learning levels, while word length has a downward sloping relationship with the elementary level only. This correlation can also be applied to materials evaluation focused on the arrangement of learning levels and chapters with respect to readability. This applicability is demonstrated by two statistical methods. First, cluster analysis is the assignment of a set of texts by level or by chapter into subsets sharing similar readability scales to evaluate materials arrangements in accordance with readability scales. Second, regression model with a breakpoint (Baayen 2008) divides a linear relation of readability scales into two models having a different slope so that we can explore a trend change of readability scales in materials. (**Research Institute for Language & Information, Korea University; Department of Linguistics, Korea University**)

Key words: Text Level, Readability, Paragraph Length, Sentence Length, Word Length, Korean Learning Materials, Materials Evaluation, Statistical Method, Correlation Analysis, Linear Regression Analysis, Regression with Breakpoint. Cluster Analysis

---

* 이 논문에서 발견되는 문제점을 지적하고, 개선을 위한 조언을 주신 익명의 심사자들께 감사한다.
** 주저자 겸 교신저자: 홍정하

1. 서론

가독성(readability)은 텍스트의 난이도를 나타내는 통계적 척도로, 텍스트 수준을 평가하는 기준으로 활용되고 있다(Flesch 1946, Fry 1977, Stenner 외 1988, 최인숙 2005 등). 일반적으로 가독성 지표는 문장, 어휘, 음절의 빈도나 길이, 또는 어휘 수준과 관련되며, 이에 기반하여 측정된 가독성 척도는 텍스트의 독자 수준 평가, 또는 독자의 읽기 능력 수준 평가에 활용되고 있다. 실제로 가독성 척도는 미국 국방성의 문서(Si · Callan 2001), 워드프로세서 내장 문서의 텍스트 수준을 설정하는 기준으로 사용되고 있으며, 특히, 교육적 목적의 읽기 능력 평가 및 텍스트 수준 평가 분야에서 활발하게 활용되고 있다. 미국에서 발행되는 교과서, 학습용 도서의 텍스트 수준 및 미국 표준 학력 평가 시험의 읽기 교과 지문은 Lexile(http://lexile.com)의 가독성 척도에 따라 평가되고 있으며, 학생의 읽기 능력을 평가하여 읽기 수준에 적합한 도서를 추천하는 Accelated Reader(http://www.renlearn.com)와 같은 교육용 독서 관리 소프트웨어에서도 가독성 척도에 기반하여 텍스트 수준을 평가하고 있다.

본 연구는 가독성 척도의 이러한 교육적 유용성을 주목하고, 외국인을 위한 한국어 학습 교재(이하 한국어 학습 교재)를 대상으로 텍스트 수준과 가독성 척도의 통계적 상관성 검증 및 이를 활용한 통계적 교재 구성 평가 방법론을 논의하는 것이 목적이다. 이를 위해 <가나다 Korean for Foreigners>[1] 초급 · 중급 · 고급 단계 총 6권, 총 171 단원의 본문 텍스트를 연구 대상 자료로 하며, 평균 문단 길이(문장수/문단수), 평균 문장 길이(어절수/문장수), 평균 어절 길이(문자수/어절수)를 관찰 대상 가독성 지표로 하여 크게 두 가지 측면에 대해 논의한다.

첫째, 한국어 학습 교재의 학습 단계별 텍스트 수준과 가독성 지표의 통계적 상관성을 검증하기 위해 학습 단계에 따른 가독성 지표 분포의 변화를 관찰하고, 상관 분석(correlation analysis), 단순 선형 회귀 분석(simple linear regression analysis), 다중 선형 회귀 분석(multiple linear regression)을 이용하여 통계적 상관성 검증 및 가독성 공식 구성을 논의한다. 둘째, 가독성 척도에 기반한 통계적 교재 구성 평가 방법론을 논의하기 위해 군집 분석(cluster analysis)을 이용하여 학습 단계 및 학습 단원 배치의 적절성을 평가하고, 구분점(breakpoint)을 포함한 선형 회귀 모형(Baayen 2008)을 활용하여

1) 가나다한국어학원 교재 연구부, 가나다 Korean for Foreigners, 서울: 랭기지플러스.

교재에 나타나는 가독성 지표의 선형적 추세 변화 구간 및 그 특성을 평가한다.

2. 선행 연구와 문제

대표적인 가독성 연구로는 Flesh Reading Ease 테스트(Flesh 1946)와 Flesch - Kincaid Grade Level 테스트(Kinkade 외 1975)가 있으며, 이 두 가독성 테스트는 평균 문장 길이(어휘수/문장수)와 어휘의 평균 음절 길이(음절수/어휘수)에 기반하여 가독성 척도를 측정한다. Flesh Reading Ease 공식 (1)은 점수가 낮을수록 높은 텍스트 난이도를 의미하며, 90.0-100.0의 점수는 11세 학생에게, 60.0-70.0의 점수는 13세-15세 학생에게, 0.0-30.0의 점수는 대학생에게 적합한 텍스트 수준으로 평가된다. Flesch-Kincaid Grade Level 공식 (2)의 점수는 미국의 학년 수준을 나타내며, 예를 들어, 점수가 8.2로 산출되었다면, 미국의 8 학년(13세-14세) 학생에게 적합한 텍스트 수준으로 평가된다.

(1) Flesh Reading Ease 공식

$$\text{텍스트 수준 점수} = 206.835 - (1.015 \times \text{평균 문장 길이}) - (84.6 \times \text{어휘 평균 음절 길이})$$

(2) Flesch - Kincaid Grade Level 공식

$$\text{텍스트 수준 점수} = (0.39 \times \text{평균 문장 길이}) + (11.8 \times \text{어휘 평균 음절 길이}) - 15.59$$

(1)과 (2)와 같이 가독성 척도는 평균 문장 길이, 어휘의 평균 음절 길이와 같은 가독성 지표로 구성된 선형 회귀 공식을 통해 산출되며, 그 값은 텍스트 수준 점수를 나타낸다. 일반적으로 가독성 연구에서 사용되는 가독성 지표는 언어 표현의 빈도수, 언어 표현의 길이, 또는 어휘 수준으로 구성되며, <표 1>은 가독성 연구별 가독성 지표 및 평가/연구 대상 텍스트 수준이다. 대체로 평균 문장 길이를 기본 가독성 지표로 사용하고 있으며, 영어 텍스트에 대해서는 어휘의 평균 음절 길이 또는 친숙성에 따른 어휘 사용을, 한국어 텍스트에 대해서는 문단 길이, 이형 어절수, 어휘 분류 등을 고려하고 있다.

<표 1> 가독성 연구: 가독성 지표 및 평가 대상 텍스트 수준

가독성 연구	가독성 지표	평가/연구 대상
Flesh Reading Ease (Flesh 1946)	- 평균 문장 길이 - 어휘의 평균 음절 길이	중학교-대학교
Flesch–Kincaid Grade Level (Kinkade 외 1975)	- 평균 문장 길이 - 어휘의 평균 어절 음절 길이	초등학교-대학교
Gunning-FOG 공식 (Gunning 1952)	- 평균 문장 길이 - 3 음절 이상 어절 비율	초등학교-고등학교
SMOG 공식 (Mclaughlin 1969)	- 30 문장당 3음절 어휘수	초등학교-고등학교
Forecast 공식 (Sticht 1973)	- 100 어휘당 1음절 어휘수	미 육군 기술 편람
Spache 공식 (Spache 1953)	- 평균 문장 길이 - 미친숙성 어휘 비율	초등학교 저학년
Dale-Chall 공식 (Dale · Chall 1948)	- 평균 문장 길이 - 미친숙성 어휘 비율	초등학교 고학년-중학교
Lexile (Stenner 외 1988)	- 평균 문장 길이 - American Heritage Intermediate Corpus 평균 어휘 빈도수 비교	초등학교-고등학교
최인숙(2005)	- 평균 문장 길이 - 어절수 - 문단수 - 문장수 - 평균 어절 길이	초등학교-고등학교
	- 평균 문단 길이 - 문단수	중학교-고등학교
	- 이형 어절수 - 신어절 출현 비율	초등학교
전정재(2001)	- 평균 문장 길이	초등학교-고등학교
심재홍(1991)	- 평균 문장 길이 - 한자어 비율 - 함축어 비율 - 지시어 비율 - 인칭 명사 비율	고등학교

가독성 연구	가독성 지표	평가/연구 대상
	- 대화 문장 비율	
최재완(1995)	- 전문용어 - 한자 - 한자어 - 외래어 - 외국문자 표기 외래어 - 숫자 - 약어 - 단어 의미 - 문장 구조 - 평균 문장 길이	대학교

그런데 한국어 텍스트에 대한 가독성 연구에서 어휘 수준을 고려한 접근은 텍스트 처리에 있어서 정확성 확보 및 고비용의 문제가 있다. 가독성 측정은 컴퓨터를 이용한 텍스트의 자동 처리를 필수적으로 수행해야 하나, 심재홍(1991), 최재완(1995)와 같이 한자어, 함축어, 지시어, 인칭 명사, 전문용어, 한자어, 외래어의 분석은 어휘 목록 작성 및 정밀한 자연언어처리 기술이 수반되어야 한다는 단점이 있다. 그래서 한국어 텍스트를 대상으로 어휘 수준과 관련한 가독성을 측정하기 위해서는 형태소 분석 등의 정확성 확보 및 분석 시스템 개발의 고비용 문제를 야기한다고 할 수 있다.[2] 반면, 언어 표현의 빈도수 또는 길이는 간단하게 자동 처리할 수 있다는 장점이 있다.[3]

이 밖에도 가독성과 관련한 기존 연구의 목적이 대체로 텍스트 수준 측정을 위한 가독성 지표 및 공식 개발에 초점을 두고 있어서, 텍스트의 독자 수준 평가, 또는 독자의 읽기 능력 수준 평가 외에는 다양한 응용 방법론이 제시되고 있지 못하다. 특히, 한국어 텍스트에 대한 연구는 아직까지 이론적 차원의 논의에 그치고 있어, 이에 대한 다양한 활용 방법론 논의가 필요하다 하겠다.

---

2) 형태소 분석의 정확성 확보 및 어휘 수준 관련 목록 구축 등의 작업이 수반된다면, 이와 관련한 가독성 논의 또한 중요하다고 할 수 있다. 이에 대한 논의는 향후 과제로 남겨 두기로 한다.

3) 물론 최인숙(2005)에서 제시한 이형 어절수, 신어절 출현 비율은 컴퓨터를 활용하여 비교적 간단하게 처리할 수 있으나, 본 연구의 관찰 대상 텍스트의 단원별 분량이 크지 않아 이러한 지표를 상정하기에 어려움이 있다. 이에 대한 본 논문의 한계에 대해서는 3.2절 참조. 또한 최인숙(2005)은 이형 어절수, 신어절 출현 비율을 초등학교 교과서에서만 통계적으로 유효할 뿐, 초중고등학교 교과서를 통합적으로 비교하거나 중고등학교 교과서만을 대상으로 한 가독성 측정에서 통계적으로 유의미하지 않은 가독성 지표로 제시하고 있다. 이러한 측면에서 이형 어절수, 신어절 출현 비율은 한국어 텍스트에 대해 보편적으로 적용하기 어려운 가독성 지표일 가능성이 있다.

한편, 가독성은 이미 외국의 교재 평가 기준으로 설정되어 있으며, 이를 객관적으로 평가할 수 있는 가독성 척도가 실제로 활용되고 있다. Zenger · Zenger(1976), Warming · Barber(1980)에서는 교재 평가 기준으로 가독성을 제시하고 있으며, 이미 미국 교과서의 평가 기준에 가독성이 포함되어 있다(진재관 외 2009). 특히, 미국에서 발행되는 초등학교에서부터 고등학교까지의 교재와 추천 도서, 그리고 미국의 표준 학력 평가 시험인 Stanford Achievement Test, Terra Nova, California Achievement Test 등의 읽기 교과 지문들은 Lexile(Stenner 외 1988)의 가독성 척도에 의해 평가되어 제시되고 있다.[4)]

다행히 최근 국내에서도 교과서 평가 기준 중 하나로 가독성이 고려되고 있다. 한국교육과정평가원에서 발간한 교과서 평가 기준(안) (3)과 교과서 평가 기준(최종안) (4)를 보면, 가독성을 평가 항목으로 제시하고 있다. 그러나 가독성을 문장의 문법성과 어휘 수준의 문제로 인식하는 측면이 있어, 컴퓨터를 이용한 객관적 평가보다는 평가자에 의한 주관적 평가에 의존할 가능성이 있다. 즉, 문장의 문법성은 현재의 한국어 자연언어처리 기술 수준에서 정확한 분석이 어렵고, 어휘 수준의 분석은 앞서 언급한 것처럼 고비용의 문제가 있으므로 실제로 이 기준에 따라 컴퓨터를 활용한 가독성 평가는 힘들다고 할 수 있다.

(3) 교과서 평가 기준(안)에서 가독성 평가 항목(진재관 외 2008)
- 문장이 명료하고 이해하기 쉬운가?
- 교과서의 어휘는 학습자 수준에 적절한가?

(4) 교과서 평가 기준(최종안)에서 가독성 평가 항목(진재관 외 2009)
- 문장이 명료하며, 어법에 맞는가?

물론 아직까지 한국어 학습 분야에서는 가독성에 대한 논의가 일부의 연구에 지나지 않지만, 국내의 교과서 평가 기준에서 제기되는 문제가 한국어 학습 분야에서도 동일하게 관찰된다. 국립국어원(2003)은 한국어 학습 단계별 어휘 수준을 분류하고 있으나, 이에 기반하여 실제 텍스트를 대상으로 가독성을 평가하기 위해서는 고비용의 문제가 있다고 할 수 있다. 또한 이해영(2001)은 한국어 학습 교재의 평가 기준으로 “읽기 자료의 길이와 난이도가 숙달도에 맞는가?”를 제안하고 있으나, 객관적 척도 제시 없이 분석자의 주관적 판단에

4) Lexile에 대한 자세한 내용은 http://lexile.com 참조.

의존한 평가 방법을 논의하고 있을 뿐이다.

특히, 한국어 학습 교재는 외국인 학습자를 대상으로 하기 때문에 학습 단계별, 그리고 학습 단원별 텍스트 난이도의 적절성이 모국어 화자를 대상으로 한 텍스트에서보다 더욱 중요하다고 할 수 있다. 이에 대한 근거 중 하나는 모국어 화자와 외국인 학습자의 언어 학습 과정에 차이가 있다는 것이다. 언어 학습 과정에 있어서 모국어 화자는 텍스트를 통해 언어를 학습하기 이전인 취학 전에 이미 대부분의 문법 지식을 습득하는 반면, 외국인 학습자는 문법 지식을 습득해 나가면서 텍스트를 접한다는 측면에서 가독성과 관련한 모국어 화자와 외국인 학습자의 특성은 구분될 수밖에 없다(Heilman 외 2007).[5] 문장 길이에 문법적 난이도가 일정 부분 반영되어 있다는 점을 고려한다면 외국인 학습자를 대상으로 하는 한국어 학습 교재의 적절한 가독성은 교재 평가의 중요한 요소라 할 수 있다. 또한 이러한 평가에서 보다 객관적인 기준을 적용하기 위해서는 컴퓨터의 활용은 필수적이라 할 수 있다.

이러한 측면에서 본 논문은 컴퓨터를 이용하여 비교적 용이하게 처리가 가능한 언어 표현의 길이를 가독성 지표로 설정하고, 한국어 학습 교재를 대상으로 가독성 지표와 텍스트 수준의 상관성에 대한 통계적 검증 및 이를 활용한 통계적 교재 구성 평가 방법론을 논의한다. 이러한 접근을 통해 첫째, 컴퓨터를 활용한 가독성 측정의 효용성 및, 둘째, 기존 가독성 연구에서 많이 다루어지지 않았던 교재 구성에 관한 평가 방법론을 확인할 수 있을 것이다.

3. 연구 방법

3.1. 연구 대상 자료와 처리

본 연구의 대상 자료인 <가나다 Korean for Foreigners>는 외국에서 일반적으로 많이 사용되는 교재일 뿐만 아니라, 국내에서도 자체 교재가 없는 학원이나 개인 학습자들이 일반적으로 선호하는 교재이다. 이 교재는 초급-1 단계 25 단원, 초급-2 단계 30 단원, 중급-1 단계 30 단원, 중급-2 단계 30 단원, 고급-1 단계 28 단원, 고급-2 단계 28 단원의 총 6권, 총 171 단원의 본문 텍스트로 구성되어 있으며, 대부분은 대화체 텍스트이다. 이 중 고급 단계의 8개 단원만이 문어 텍스트이며, 일부 단원(특히, 고급-2 단계)은 대화 상황을

5) 모국어 화자와 외국인 학습자의 이러한 차이는 가독성 지표 및 공식에 차이가 있을 수 있다. 이에 대한 논의는 향후 연구로 남기기로 한다.

기술하는 문어 텍스트가 혼용되어 있는 대화체 텍스트이다. 그러나 총 171 단원 중 문어 텍스트의 비중이 적으므로, 본 연구에서는 대화체 텍스트와 문어 텍스트를 구분하여 않는다. 또한 단원 제목, 대화체 텍스트에서 대화자 표지, 한국어 텍스트를 번역한 영문 텍스트, 영문 병기 표기, 문장 부호, 공백 문자는 가독성 산출 대상에서 제외한다. 따라서 가독성 측정 대상인 문자열은 본문 텍스트에 나타나 있는 한글 문자열 및 숫자이다.

한편, 본 연구의 통계적 검증 및 활용은 한국어 학습 교재의 텍스트 난이도가 학습 단계 및 단원의 순서에 따라 순차적으로 높아진다는, 그리고 높아져야 한다는 인식에 기초한다. 이러한 가정을 통계적으로 검증하고 활용하기 위해 학습 단계 및 학습 단원에 따라 텍스트 수준 코드를 구분하여 사용한다. (5)와 (6)은 각각 학습 단계별 텍스트 수준 코드와 학습 단원별 텍스트 수준 코드이다. (5-가)는 학습 단계의 대분류 코드를, (5-나)는 학습 단계의 소분류 코드를 나타내며, (6-가)는 소분류 학습 단계별 단원 구분 코드를, (6-나)는 초급부터 고급까지의 단원 순서 코드를 나타낸다. 또한 (5-다)와 (6-나)는 학습 단계 및 학습 단원의 텍스트 수준 점수로 활용된다.

(5) 학습 단계별 텍스트 수준 코드
  가. 초급, 중급, 고급: 각각 A, B, C
  나. 초급-1, 초급-2, 중급-1, 중급-2, 고급-1, 고급-2:
    각각 A1, A2, B1, B2, C1, C2
  다. 학습 단계별 텍스트 수준 점수: 1, 2, 3, 4, 5, 6

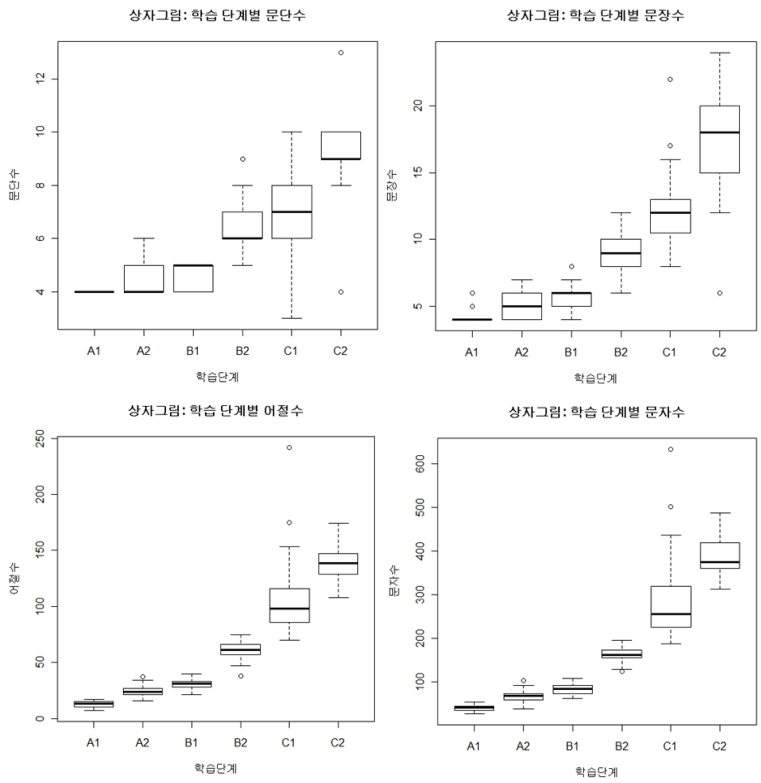
(6) 학습 단원별 텍스트 수준 코드
  가. 초급-1 1 단원, 초급-1 2 단원, …, 고급-2 27 단원, 고급-2 28 단원:
    각각 A1-1, A1-2, …, C2-27, C2-28
  나. 1 단원, 2 단원, …, 170 단원, 171 단원:
    (학습 단원별 텍스트 수준 점수)
    각각 1, 2, …, 170, 171

3.2. 관찰 대상 가독성 지표

본 연구 대상 자료는 학습 단계 순서에 따라 텍스트의 분량이 점차 증가하는 추세를 보인다. 즉, 각 단원의 텍스트 분량은 대체로 초급-1 단계(A1)에서 가장 적으며, 학습 단계 순서에 따라 점차 증가하여 고급-2 단계(C2)에서

텍스트 분량이 가장 많다. 이는 텍스트 분량의 영향을 많이 받는 문단수, 문장수, 어절수, 문자수의 순차적 증가 추세에서도 확인할 수 있다.

<그림 1>은 학습 단계별 문단수, 문장수, 어절수, 문자수의 분포 범위를 나타낸 상자 그림(box plot)이다. 상자 그림은 관측치를 크기에 따라 배열하여 전체 관측치를 사등분하는 위치의 값인 사분위수 범위(Interquartile Range)를, 즉, 최소값(0/4분위수), 1/4분위수, 중앙값(2/4분위수), 3/4분위수, 최대값(4/4분위수)를 통해 관측 자료의 산포도를 나타낸다. 예를 들어, <그림-1>에서 각 상자의 가로 경계선은 아래에서부터 1/4분위수, 중앙값(굵은선), 3/4분위수를 나타내며, 상자 위 · 아래에 점선으로 연결되어 있는 가로 경계선은 최대값과 최소값을, 최대값 또는 최소값 위 · 아래의 점은 특이값(outlier)을 표시한다. 이 상자 그림들을 통해 <그림 1>에서 텍스트 분량과 관련된 문단수, 문장수, 어절수, 문자수는 모두 증가하는 분포 추세를 보인다. 물론 이러한 증가 추세를 개별 지표의 분포적 증가 추세로 볼 수도 있겠으나, 이보다 텍스트 분량의 증가에 따라 이들 지표의 분포적 증가로 보는 것이 적절해 보인다.



<그림 1> 학습 단계별 언어 표현 빈도수 분포의 상자 그림

그러나 이러한 특성은 가독성 측정을 위한 표본의 표준 분량 설정을 어렵게 만드는 원인이 된다. 일반적으로 언어 표현의 빈도수와 관련한 가독성 연구에서는 관찰 대상 텍스트 분량을 일정 규모로 표준화하고 있다. 3 음절 어휘의 빈도를 다루는 SMOG 공식(Mclaughlin 1969)은 30 문장으로, 1 음절 어휘의 빈도를 계산하는 Forecast 공식(Sticht 1973)은 100 어휘로, 문단, 문장, 어절 등의 빈도를 측정하는 최인숙(2005)은 2,000 문자로 텍스트 분량을 표준화하고 있다.

이에 비해 본 연구 대상 자료는 학습 단계에 따라 텍스트 분량의 편차가 크고, 단원별 텍스트 분량이 적어 표본 텍스트의 분량을 표준화하기 어려운 측면이 있다. 대체로 초급-1 단계의 단원별 분량은 4 문장, 13 어절, 40 문자, 중급-2 단계의 단원별 분량은 9 문장, 61 어절, 163 문자, 고급-2 단계의 단원별 분량은 17 문장, 138 어절, 386 문자 정도에 불과하다. 이러한 단원별 텍스트 분량은 기존 연구의 표본 분량에 크게 못 미치는 수준이다. 물론 학습 단계별 텍스트를 통합하여 표본을 추출할 수도 있겠으나, 학습 단계 및 학습 단원에 따른 가독성 지표의 특성을 관찰하고자 하는 본 논문의 목적에 부합하지 않는다. 또한 코퍼스의 일반적인 분량 단위인 어절 단위로 텍스트 분량을 표준화할 경우, 문단 및 문장이 부적절하게 분할되는 문제도 있다.

그래서 본 논문에서는 텍스트 분량의 영향을 많이 받는 언어 표현의 빈도수, 즉, 문단수, 문장수, 어절수, 문자수는 가독성 측정 지표에서 제외한다. 이에 비해 언어 표현의 길이 (7)은 평균치로 산출되기 때문에 텍스트 분량의 영향을 비교적 적게 받으므로, 본 연구에서는 (7)을 관찰 대상 가독성 지표로 한다.

(7) 관찰 가독성 지표
  가. 평균 문단 길이
  나. 평균 문장 길이
  다. 평균 어절 길이

3.3. 통계 분석

이 논문에서 텍스트 수준에 따른 가독성 지표의 상관성 검증 및 교재 구성 평가를 위해 (8)의 세 가지 통계 분석 방법, 즉, 상관 분석, 선형 회귀 분석, 군집 분석을 사용한다.[6]

6) 이 논문의 통계 분석은 R 통계 패키지를 이용한다. R 통계 패키지는 공개용 소프트웨어로 통계 처리 외에도 텍스트 처리, 프로그래밍 기능을 지원하고 있다(http://www.r-project.org/).

(8) 통계 분석 방법

가. 상관 분석(correlation analysis)

나. 선형 회귀 분석(linear regression analysis)

다. 군집 분석(cluster analysis)

첫째, 상관 분석은 두 변수 사이의 선형 관계에 대한 상관성 및 방향성을 분석하는 통계 기법으로 두 변수의 상호 관계를 나타낸다. 상관 분석은 - 1과 + 1 사이의 상관계수를 산출하며, 그 값이 0에 근접하면 두 변수 사이의 상관성이 없는 것으로, + 1 또는 - 1에 근접하면 상관성이 큰 것으로 해석된다. 또한 + 값의 상관계수는 두 변수의 비례적 상관성을, - 값의 상관계수는 반비례적 상관성을 나타낸다. 특히, 상관계수의 값이 - 1부터 + 1까지의 동일 척도로 제시되므로 다양한 변수의 상관성 크기를 비교할 수 있다는 장점이 있다. (9)는 상관계수 절대값 범위에 따른 통계적 판별 기준을 Cohen(1988)에서 제시한 것이다. 예를 들어, 상관 분석 결과로 상관계수의 값이 + 0.90 또는 - 0.90이 나왔다면, (9)에 따라 두 변수는 높은 상관성을 갖는다고 판별할 수 있다. 4.2 절에서는 가독성 연구에서 일반적으로 사용되는 피어슨 적률 상관계수(Pearson's product-moment correlation)를 활용하여 텍스트 수준과 가독성 지표의 상관성을 분석하고 (9)를 활용하여 상관도를 평가한다.

(9) 상관계수(절대값)의 판별(Cohen 1988)[7]

0.00 - 0.09 : 상관관계가 없다(none)

0.10 - 0.30 : 낮은 상관관계(small)

0.30 - 0.50 : 비교적 높은 상관관계(medium)

0.50 - 1.00 : 높은 상관관계(large)

둘째, 선형 회귀 분석은 하나 이상의 독립 변수와 하나의 종속 변수로 구성되며, 종속 변수에 대한 독립 변수의 영향력, 즉, 종속변수에 대한 독립 변수의 인과 관계를 선형 관계식으로 추정하고 분석하는 통계 기법이다. 상관 분석과의 주요 차이점이라면 상관 분석은 독립 변수와 종속 변수의 구분 없이 쌍방간의 상관성을, 선형 회귀 분석은 종속 변수에 대한 독립 변수의 영향력을 나타낸다는 것이다. 또한 상관 분석의 상관계수의 범위는 동일한 척도이지만, 선형

7) 상관계수 값의 범위와 이에 따른 통계적 해석에 절대적 기준이 있는 것은 아니어서 통용되는 기준 또한 다양하나, 본 연구에서는 Cohen(1988)에서 제시한 (9)의 기준에 따라 상관도를 판별한다.

회귀 분석의 회귀계수의 범위는 제한되어 있지 않다는 것이다.

(10) 선형 회귀 공식 형식
가. 단순 선형 회귀: $Y = \alpha + \beta X$
나. 다중 선형 회귀: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$

일반적으로 가독성 공식은 선형 회귀식으로 구성되는데, 이는 함수 형태 중 가장 단순하고 다루기 쉬운 직선의 관계식으로 산출되기 때문이다. 단순 선형 회귀 모형은 하나의 종속 변수와 하나의 독립 변수로 구성되며, 공식의 형식은 (10-가)와 같다. 여기서 Y는 선형회귀 모형에 의해 산출되는 종속 변수의 값을, X는 종속 변수에 대해 영향력을 미치는 독립 변수의 값을, β는 선형 회귀 모형에서 독립 변수가 갖는 영향력의 크기인 직선의 기울기를, α는 선형 회귀 모형의 기본 값을 나타내는 직선의 절편을 나타낸다. 선형 회귀 모형에서는 β로 표시되는 회귀계수가 가장 중요하며, 이를 통해 독립 변수와 종속 변수의 선형 관계를 평가한다. 또한 하나의 종속 변수에 대해 영향을 미치는 독립 변수가 여러 개일 경우 다중 선형 회귀 모형을 사용하며, 그 공식의 형식은 (10-나)와 같다. 4.2 절에서는 단순 선형 회귀 분석을 통해 학습 단계와 가독성 지표의 상관성을 평가하고, 다중 선형 회귀 분석을 통해 한국어 학습 교재의 텍스트 수준을 평가할 수 있는 가독성 공식을 제시한다.

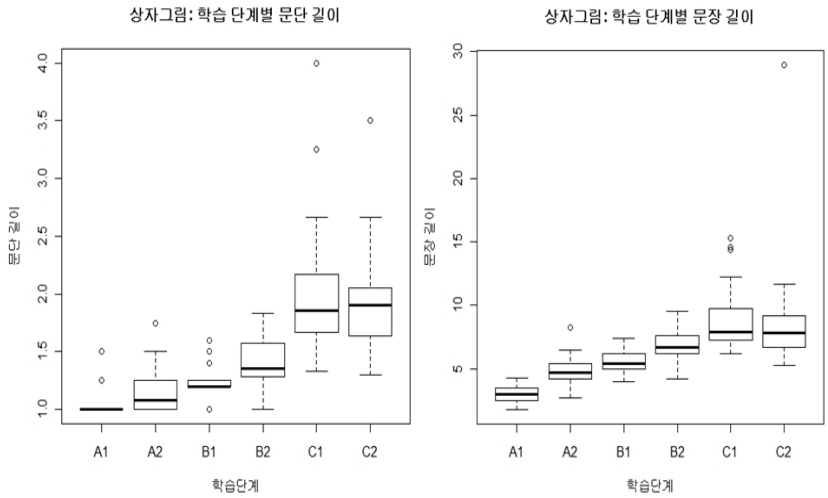
셋째, 군집 분석은 변수들의 유사성과 비유사성에 기반하여 변수들을 군집화하는 통계 기법으로 계층적 군집 분석 기법이 대표적이다. 본 논문에서는 계층적 군집 분석의 일반적인 통계 기법인 응집적 군집 분석(agglomerative clustering)을 활용한다. 응집적 군집 분석은 상향식(bottom-up) 방식으로 모든 변수들을 유사성과 비유사성에 따라 순차적으로 군집화하여 최종적으로 하나의 군집으로 분류하며, 그 분류 결과는 수형 구조도(dendrogram)로 표시된다. 군집 분석은 5 절에서 군집화된 학습 단계 및 학습 단원의 텍스트를 통해 교재 배치의 적절성을 평가하는 통계적 방법으로 활용된다.

4. 가독성 검증

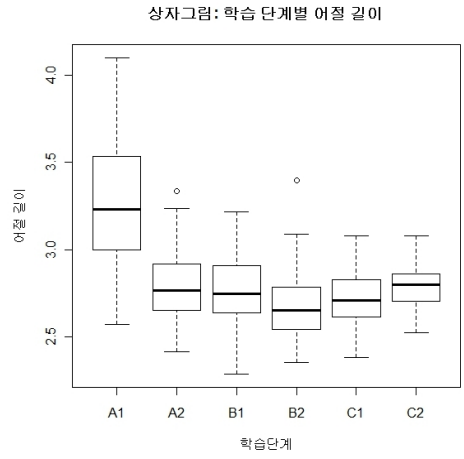
4.1. 학습 단계별 분포

<그림 2>와 <그림 3>은 학습 단계별 텍스트의 평균 문단 길이(이하 문단

길이), 평균 문장 길이(이하 문장 길이), 평균 어절 길이(이하 어절 길이)를 상자 그림으로 제시한 분포이며, 각 상자의 가로 경계선은 아래에서부터 1/4분위수, 중앙값(굵은선), 3/4분위수를 나타내며, 상자 위 · 아래에 점선으로 연결되어 있는 가로 경계선은 최대값과 최소값을, 최대값 또는 최소값 위 · 아래의 점은 특이값(outlier)을 표시한다. 상자 그림을 통해 학습 단계의 증가에 따라 <그림 2>에서 문단 길이와 문장 길이의 증가 추세를, <그림 3>에서 어절 길이의 감소 추세를 관찰할 수 있다.



<그림 2> 학습 단계별 문단 길이/문장 길이의 상자 그림



<그림 3> 학습 단계별 어절 길이의 상자 그림

문단 길이 및 문장 길이의 이러한 증가 추세는 학습 단계에 따른 담화 및 구문의 복잡성을 나타내는 것으로 파악할 수 있다. 즉, 텍스트 난이도가 높아짐에 따라 담화적 복잡성을 나타내는 문단 길이와 구문적 복잡성을 나타내는 문장 길이가 점차 길어지는 것으로 이해할 수 있다. 그런데 어절 길이의 감소 추세는 영어에서 의미적 복잡성을 반영하는 어휘 길이의 증가 추세와 대조된다. 이는 한국어의 어휘 길이가 영어에서처럼 다양하지 않을 뿐만 아니라, 어절이 어휘에 문법 형태소가 결합되어 구성된다는 점에서 영어의 어휘 길이 추세와는 다른 양상을 보이는 것으로 판단된다.

<표 2> 국립국어원(2003)의 학습 수준별 어휘 길이

학습 수준	총문자수	총어휘수	평균 어휘 길이
초급	2,052	894	2.295
중급	5,246	2,028	2.587
고급	7,522	2,788	2.698
합계	14,820	5,710	2.595

<표 2>는 국립국어원(2003)에서 제시한 한국어 학습자용 학습 수준별 어휘 목록을 토대로 측정한 학습 수준별 어휘 길이이다. 물론 한국어의 어휘는 영어 어휘처럼 길이의 편차가 크지 않아 학습 수준별 어휘 길이에 차이가 크지 않지만, 영어와 유사하게 학습 수준이 높아질수록 어휘 길이는 점차 증가하는 추세를 보인다. 즉, 한국어에서도 어휘 길이가 어휘 의미의 복잡성을 반영할 가능성이 있다고 할 수 있다. 그러나 이는 학습 단계에 따른 어절 길이의 분포 추세 <그림 3>과 반대의 결과로, 어절 길이가 어휘 의미의 복잡성을 반영한다고 보기에 어려움이 있다.[8)]

오히려 어절 길이의 감소 추세는 짧은 문장(또는 텍스트 분량)에서 특정 어절이 차지하는 분포적 비중에 의해 발생되는 것으로 보인다. 어절 길이 측정은 문장 또는 텍스트의 [문자수 ÷ 어절수]를 통해 산출되므로, 어절 길이 측정에서 어절수가 많지 않은 경우, 즉, 문장 또는 텍스트 길이가 짧은 경우에 특정 어절의 분포적 영향력이 큰 반면, 어절수가 많은 경우, 즉, 문장 또는 텍스트 길이가 긴 경우에 특정 어절의 분포적 영향력이 약화되기 마련이다.

예를 들어, (11)은 초급 단계, 중급 단계에서 발췌한 예문인데, 이 문장들의

8) 물론 <표 2>만으로 한국어의 어휘 난이도와 어휘 길이의 관련성, 그리고 어절 길이와 어휘 길이의 관련성을 평가하기는 어렵다. 이러한 상관성은 통계적 방법론을 통한 검증도 필요하지만, 어절에 대한 정확한 형태소 분석을 수행해야 한다. 본 연구에서는 형태소 단위의 가독성을 다루지 않으므로 향후 연구로 남기기로 한다.

어절수, 문자수, 어절 길이를 측정한 수치는 (12)와 같다. 초급 단계 문장 (11-가)와 중급 단계 문장 (11-나)에서 가장 긴 어절 "읽으십니까"와 "준비하도록"을 제외한 상태에서 어절 길이를 산출하면, 초급 단계 문장은 2.0, 중급 단계 문장은 2.25으로 중급 단계의 어절 길이가 더 길게 측정된다. 그러나 각 문장에서 가장 길면서 동일하게 5 어절인 서술어를 포함하여 어절 길이를 산출하면 각각 2.75 (12-가)와 2.556 (12-나)로 오히려 초급 단계의 어절 길이가 더 길게 측정된다. 이처럼 짧은 문장에서는 특정 어절의 길이가 어절 길이 측정에 미치는 영향력이 크므로, 중급 및 고급 단계에 비해 문장 길이가 짧은 초급 단계의 어절 길이는 길어지는 경향을 보인다. 특히, 한국어 학습 교재에서 주로 사용되는 경어체 서술어는 다른 어절보다 어절 길이가 더 길어지는 경향이 있으므로, 짧은 문장에 대한 어절 길이 측정에서 경어체 서술어의 분포적 영향력은 더욱 확대된다고 할 수 있다.[9)]

(11) 학습 단계별 문장의 예
가. 초급: 한국 신문을 안 읽으십니까?
나. 중급: 음식은 너무 많이 하지 말고 먹을 만큼만 준비하도록 해요.

(12) 예문 (8)의 어절수, 문자수, 어절 길이
가. 초급: 어절수 4, 문자수 11, 어절 길이 2.750
나. 중급: 어절수 9, 문자수 23, 어절 길이 2.556

한편, 한국어 학습 교재에 나타나는 문단 길이 및 문장 길이의 증가 추세, 어절 길이의 감소 추세는 초중고 교과서에서도 동일하게 관찰된다. 최인숙(2005)은 초중고 교과서를 대상으로 한 가독성 측정에서 텍스트 수준에 따른 문장 길이의 증가 추세 및 어절 길이의 감소 추세를, 중고 교과서를 대상으로 한 가독성 측정에서 문단 길이의 증가 추세를 통계적으로 유의미한 상관성으로 분석하고 있다. 이러한 측면에서 문단 길이, 문장 길이, 어절 길이를 한국어 학습 교재뿐만 아니라, 한국어 텍스트에 보편적으로 관찰되는 가독성 지표로 고려할 수 있으며, 그 선형적인 특성 또한 유사한 것으로 파악할 수 있다.

그러나 가독성 지표로서 문단 길이, 문장 길이, 어절 길이를 보다 객관적으로 논의하기 위해서는 통계적 검증을 통해 텍스트 수준과의 상관성을 검토할 필요가 있다. 이를 위해 4.2 절에서는 상관 분석과 선형 회귀 분석을 통해

9) 초중고 교과서를 대상으로 어절 길이의 감소 추세를 관찰하고 있는 최인숙(2005)에서도 어절 길이의 감소 추세를 초등학교 텍스트의 빈번한 경어체 사용으로 인한 현상으로 보고 있다.

통계적 상관성을 검증하고, 이들 가독성 지표로 구성되는 가독성 공식의 통계적 유의미성을 논의한다.

4.2. 통계적 상관성

<표 3>은 학습 단계의 텍스트 수준 점수와 가독성 지표 사이의 상관 분석 결과이다. 3.3 절의 상관계수의 판별 (9)에 따라 문단 길이, 문장 길이는 “높은 상관관계”로, 어절 길이는 “비교적 높은 상관관계”로 판별될 수 있다. 또한 학습 단계의 텍스트 수준에 대해 양의 상관계수 값을 갖는 문단 길이와 문장 길이는 비례적 상관관계를, 음의 상관계수 값을 갖는 어절 길이는 반비례적 상관관계를 나타낸다. 특히, 모든 가독성 지표의 p-value가 유의 수준 0.001보다 작으므로 모든 가독성 지표가 학습 단계의 텍스트 수준과 통계적으로 유의미한 상관성을 갖는다.[10]

<표 3> 학습 단계와 가독성 지표의 상관 분석

가독성 지표	상관계수	p-value
문단 길이	0.687513	< 2.2e-16
문장 길이	0.668	< 2.2e-16
어절 길이	-0.4073	3.54e-08

그런데 상관 분석은 독립 변수와 종속 변수의 구별 없이 두 변수의 쌍방적 상관성을 분석하는 반면, 선형 회귀 분석은 독립 변수와 종속 변수가 구분되어 독립 변수의 변화량에 따른 종속 변수의 변화량을 나타낸다. 즉, 종속 변수에 대해 갖는 독립 변수의 영향력을 나타낸다. 그래서 단순 선형 회귀 분석에서는 학습 단계별 텍스트 수준을 독립 변수로도, 그리고 종속 변수로도 볼 수 있다.

가독성 연구에서는 텍스트를 구성하고 있는 가독성 지표를 기반하여 텍스트 수준을 측정하는 것이 주목적이므로, 일반적으로 후자의 방식으로 선형 회귀 모형을 구축한다.[11] 물론 이러한 접근 방식은 각 가독성 지표의 값에 따라

10) <표 3>의 p-value 수치에 포함된 로마자 e는 부동소수점(floating point)을 나타낸다. 부동소수점은 로마자 e 또는 E로 표현되며, 선행하는 숫자에 밑이 10인 지수를 곱한 수치를 표현한다. 그래서 12e+2와 12e-2로 표시된 값은 12×10^{2} = 1,200와 12×10^{-2} = 0.12를 의미한다. 보다 쉽게 설명하자면, e 또는 E 앞의 값에 대해 e 또는 E 뒤의 숫자만큼 +는 오른쪽으로, -는 왼쪽으로 소수점을 이동시키면 된다. 그래서 12e+2와 12e-2는 e 앞의 값 12에 대해 e 뒤의 숫자만큼 소수점을 각각 오른쪽으로 두 자리, 왼쪽으로 두 자리 이동시킨다.

예측 가능한 텍스트 수준, 그리고 이에 대한 개별 가독성 지표의 영향력을 보여주므로 텍스트를 구성하는 개별 가독성 지표에 기반하여 텍스트 수준을 평가하는 용도에서는 유용하다. 그러나 전자의 접근 방식은 학습 단계, 즉, 텍스트 수준의 값에 따라 예측 가능한 개별 가독성 지표의 값, 그리고 이에 대한 텍스트 수준의 영향력을 제시하므로, 교재 개발 단계에서는 유용하다고 할 수 있다. 이처럼 두 가지 방식의 선형 회귀 분석은 서로 다른 측면을 반영하므로, 본 연구에서는 두 가지 방식의 선형 관계를 모두 제시한다.[12]

<표 4> 학습 단계(독립 변수)와 가독성 지표의 단순 선형 회귀 분석

가독성 지표	절편	t-value	p-value	회귀계수	t-value	p-value
문단 길이	0.75339	11.67	< 2e-16	0.20345	12.27	< 2e-16
문장 길이	2.1679	5.509	1.34e-07	1.1758	11.635	< 2e-16
어절 길이	3.08375	63.3	< 2e-16	-0.07231	-5.78	3.54e-08

가독성 지표	F-value	p-value	R^2	조정 R^2
문단 길이	150.6	< 2.2e-16	0.4727	0.4695
문장 길이	135.4	< 2.2e-16	0.4462	0.4429
어절 길이	33.41	3.54e-08	0.1659	0.1609

<표 4>는 학습 단계의 텍스트 수준을 독립 변수로, 가독성 지표를 종속 변수로 하여 단순 선형 회귀 분석을 수행한 결과이다. 모든 가독성 지표에서 절편 및 회귀계수의 t-value에 대한 p-value가 유의 수준 0.001보다 낮으므로, 모든 가독성 지표의 절편 및 회귀계수 값이 통계적으로 유의미하다. 또한 분산 분석(analysis of variance)에 의한 F-value의 p-value도 통계적으로 유의미하므로,[13] 절편 및 회귀계수로 구성된 각 선형 회귀 모형도 통계적으로 유의미하다고 할 수 있다. 가독성 지표별 절편 및 회귀계수는 3.3 절 단순

11) 초중고 교과서를 대상으로 가독성 측정을 한 최인숙(2005)에서도 학습 수준을 의존 변수로 한 단순 선형 회귀식만을 제시하고 있을 뿐이다.

12) 익명의 심사자는 본 연구에서 사용하는 상관 분석, 선형 회귀 분석보다 둘 이상 집단의 평균을 비교하여 균일성을 평가하는 분산 분석이 본 논문에 더 적합한 통계 검증이라고 제안하고 있다. 그러나 분산 분석은 가독성 지표와 관련한 학습 단계별 텍스트의 균일성을 평가할 수 있으나, 학습 단계에 따라 가독성 지표가 증가하는지, 또는 감소하는지에 대한 통계적 방향성은 제시하지 못하므로, 통계적 방향성 포착이 중요한 본 연구에는 부적합한 통계 검증법이다.

13) 선형 회귀 분석에서 분산 분석에 의한 F-value는 단순 선형 회귀 분석에서 큰 의미가 없고, 여러 독립 변수가 포함된 다중 선형 회귀 분석에서 선형 회귀 모형의 통계적 유의미성을 판단하는 기준으로 활용된다.

선형 회귀 공식의 형식 (10-가)에서, 즉, [$Y = \alpha + \beta X$]에서 각각 α와 β에 해당되며, 이 공식을 통해 학습 단계에 따라 개별 가독성 지표의 값을 예측할 수 있다. 예를 들어 중급-1 단계의 수준 점수 3을 독립 변수 X에 대입하여 지표별 절편 및 회귀계수에 따라 산출되는 의존 변수 Y는 문단 길이에서 1.36374, 문장 길이에서 5.6953, 어절 길이에서 2.86682로 산출된다. 또한 회귀계수를 통해 개별 가독성 지표의 값을 결정하는 학습 단계의 영향력 크기를 추산할 수 있다. <표-4>에서 회귀계수는 독립 변수인 학습 단계별 텍스트 수준이 1씩 증가할 때, 각 가독성 지표 값의 증감 변동량을 의미하므로, 텍스트 수준이 1씩 증가할 때마다 문단 길이의 값은 0.20345, 문장 길이의 값은 1.1758, 어절 길이의 값은 -0.07231 변화함을 나타낸다. 이에 따라 텍스트 수준의 변화에 따라 문장 길이의 변동폭이 가장 크다고 할 수 있다.

한편, R^2은 0에서 1사이의 값을 가지며, 회귀식을 도출한 관측 자료 중에서 회귀식으로 설명 가능한 관측 자료의 분포 비율, 즉, 회귀 모형의 적합도(goodness of fit)를 나타낸다. 이 값이 1에 가까우면 선형 모형에 의해 설명 가능한 관측 자료의 분포 비율이 크며, 이에 따라 회귀 모형의 적합도가 높다는 의미이다. 그래서 <표-4>에서 R^2 값이 0.4727인 문단 길이의 자료들은 해당 선형 회귀 모형에 의해 47.27%, R^2 값이 0.4462인 문장 길이의 자료들은 44.62%, R^2 값이 0.1659인 어절 길이의 자료들은 16.59% 설명될 수 있음을 의미한다. 또한 조정 R^2(adjusted R^2)은 독립 변수의 개수가 고려되어 조정된 R^2로 다중 회귀 분석에서는 이 값을 사용한다.[14]

(13) R^2 값의 판별(Cohen 1988)[15]

0.01 : 낮은 효과(small effect)
0.09 : 중간 효과(medium effect)
0.25 : 높은 효과(large effect)

이 수치들을 Cohen(1988)에서 제시한 R^2 값의 판별 기준 (13)에 따라 구분하면 문단 길이 및 문장 길이는 “높은 효과”로, 어절 길이는 “중간 효과”로 판별할 수 있다. 이러한 결과를 통해 어절 길이의 선형 모형이 통계적으로

14) 학습 단계를 독립 변수로 하는 단순 선형 회귀 분석 <표-4>의 가독성 지표별 조정 R^2는 5.2 절에서 구분점을 포함한 선형 회귀 모형(regression with breakpoint)의 가독성 지표별 조정 R^2과 다시 비교하여 제시될 것이다.

15) 상관계수와 마찬가지로 R^2 값에 따른 통계적 해석에 절대적 기준이 있는 것이 아니다. 자연과학 분야에서는 R^2 값을 0.7 이상 요구하기도 하나, 본 연구에서는 Cohen(1988)에서 제시한 (13)의 기준에 따라 판별한다.

유의미할지라도, 어절 길이의 관측 자료들 중 83.41%는 어절 길이의 회귀식을 벗어나 분포하므로, 통계적 설명력, 또는 통계적 예측력이 비교적 약하다고 할 수 있다.

<표 5>는 학습 단계의 텍스트 수준을 종속 변수로, 가독성 지표를 독립 변수로 한 단순 선형 회귀 분석 결과이다. 문단 길이 절편에 대한 t-value의 p-value를 제외한 가독성 지표의 절편, 회귀계수, F-value는 모두 p-value가 유의 수준 0.001보다 작으므로, 통계적으로 유의미하다고 할 수 있다.[16] <표 5>의 절편 및 회귀계수를 통해 개별 가독성 지표의 값에 따라 텍스트 수준을 예측할 수 있으며, 회귀계수를 통해 텍스트 수준 값은 문장 길이에 비해 문단 길이 및 문장 길이의 영향력을 많이 받는다고 할 수 있다. 또한, <표 4>와 마찬가지로 R^2 값이 낮은 어절 길이에 대한 선형 모형은 여전히 통계적 예측력이 비교적 약하다고 할 수 있다.[17]

<표 5> 학습 단계(의존 변수)와 가독성 지표의 단순 선형 회귀 분석

가독성 지표	절편	t-value	p-value	회귀계수	t-value	p-value
문단 길이	0.1046	0.356	0.722	2.3233	12.271	<2e-16
문장 길이	1.12529	4.96	1.72e-06	0.37949	11.63	< 2e-16
어절 길이	10.0086	8.864	1.08e-15	-2.2941	-5.78	3.54e-08

가독성 지표	F-value	p-value	R^2	조정 R^2
문단 길이	150.6	< 2.2e-16	0.4727	0.4695
문장 길이	135.4	< 2.2e-16	0.4462	0.4429
어절 길이	33.41	3.54e-08	0.1659	0.1609

이상과 같이 상관 분석과 단순 회귀 분석을 통해 학습 단계별 텍스트 수준과 관찰 가독성 지표인 문단 길이, 문장 길이, 어절 길이에 대한 통계적 검증을 살펴보았다. 단순 회귀 분석에서 비록 어절 길이에 대한 R^2 값이 낮아 선형 모형으로서 통계적 예측력이 약하긴 하지만, 관찰 가독성 지표들은 모두 텍스트 난이도와 통계적 상관성을 보이고 있다. 그러나 가독성 연구에서는 텍스트 수준과 개별 지표 사이의 상관성보다 하나 이상의 가독성 지표로 구성된 선형 회귀 모형을 가독성 측정에 활용하고 있다. 이를 위해서는 가독성 지표들

16) 선형 회귀 분석에서 절편 및 회귀계수의 t-value는 해당 계수의 값이 0인지 아닌지를 판단하는 근거이다. <표 5>에서 문단 길이의 절편은 p-value가 높으므로 원점을 지나는 회귀식을 의미한다.
17) 학습 단계를 의존 변수로 하는 단순 회귀 분석 <표-5>의 가독성 지표별 조정 R^2는 <표-7>에서 다중 선형 회귀 모형의 조정 R^2과 다시 비교하여 제시될 것이다.

을 독립 변수로 하여 구성 가능한 다양한 선형 회귀 모형을 평가하고, 최적의 선형 회귀 모형을 선별하는 과정이 필요하다.

이 논문에서는 한국어 학습 교재에 대한 최적의 가독성 선형 회귀 모형을 탐색하기 위해 AIC(Akaike information criterion)를 활용한다(Akaike 1974). AIC는 제시된 변수로 구성 가능한 후보 통계 모형을 생성하고, 이 중에서 적합도 모형을 추출한다.[18] (14-가)는 AIC 수행을 위해 입력된 변수 목록이며, (14-나)는 이를 기반하여 AIC에서 생성된 후보 선형 회귀 모형을 구성하는 독립 변수이다.

(14) AIC 수행을 위한 입력 변수 및 후보 모형

가. 입력 변수

- 의존 변수: 학습 단계
- 독립 변수: 문단 길이, 문장 길이, 어절 길이

나. 후보 모형

- 문단 길이
- 문장 길이
- 어절 길이
- 문단 길이 + 문장 길이
- 문단 길이 + 어절 길이
- 문장 길이 + 어절 길이
- 문단 길이 + 문장 길이 + 어절 길이

(15) 한국어 학습 교재의 가독성 공식[19]

- $Y = 2.80763 + 1.56019X_1 + 0.2156X_2 - 1.03963X_3$
- X_1 : 문단 길이, X_2 : 문장 길이, X_3 : 어절 길이

---

18) R 통계 패키지에서 AIC는 step( ) 함수를 이용하여 실행된다.

19) 익명의 심사자는 회귀 모형의 타당성을 검증하기 위해 잔차 분석이나 오차의 정규성 검토의 필요성을 언급하고 있다. 물론 여러 종의 한국어 학습 교재 전체를 대상으로 하여 교재 평가를 위한 표준 회귀 공식 도출이 본 논문의 목적이라면 회귀 모형의 타당성 검증은 필수적이라 할 수 있다. 그러나 본 논문은 정밀한 통계적 공식 제시보다는 텍스트 난이도와 가독성 지표의 상관성을 제시하기 위해 회귀 모형을 사용하는 것이므로, 회귀 모형의 타당성 검증은 제시하지 않는다. 또한 선형회귀 모형은 일반적으로 정규 분포를 가정하는데, 본 논문에서는 이에 대한 검증을 제시하지 않는다. 설령 본 논문에서 관찰하는 관찰 자료가 정규 분포를 따르지 않는다고 하더라도, 표본 크기가 크면(일반적으로 30이상) 중심극한정리에 의해 정규분포이론을 적용할 수 있다고 알려져 있으므로 정규 분포에 대한 검증은 큰 문제가 되지 않는다.

<표 6> 가독성 적합도 선형 회귀 모형의 분석 결과

모형	계수	t-value	p-value	F-value	p-value	조정 R^2
절편	2.80763	3.174	1.79e-03			
문단 길이	1.56019	8.689	3.35e-15			
문장 길이	0.2156	6.897	1.06e-10			
어절 길이	-1.03963	-3.767	0.000229			
				101.1	< 2.2e-16	0.6398

(14-나)의 후보 모형 중에서 AIC에 의해 적합도 모형으로 선택된 회귀 모형은 [문단 길이 + 문장 길이 + 어절 길이]를 독립 변수로 하여 구성된 다중 회귀 모형으로, 이 적합도 모형의 가독성 공식은 (15)와 같다. 이 적합도 모형의 회귀 분석 결과 <표 6>에서 절편, 회귀계수, F-value에 대한 p-value가 유의 수준 0.01보다 작으므로 이 적합도 모형은 통계적으로 유의미하다.

<표 7> 후보 모형의 R^2/조정 R^2 비교

후보 모형	R^2	조정 R^2
문단 길이	0.4727	0.4695
문장 길이	0.4462	0.4429
어절 길이	0.1659	0.1609
문단 길이 + 문장 길이	0.6159	0.6113
문단 길이 + 어절 길이	0.5448	0.5393
문장 길이 + 어절 길이	0.4853	0.4791
문단 길이 + 문장 길이 + 어절 길이	0.6462	0.6398

또한 조정 R^2 값 0.6398은 Cohen(1988)의 R^2 값 판별 기준 (13)에 따라"높은 효과"로 판별될 수 있으며,[20] 학습 단계를 의존 변수로 하는 <표 5>의 단순

20) 일반적으로 단순 선형 회귀 모형에서는 R^2 값을 통해 통계적 설명력을 파악하지만, 다중 선형 회귀 모형에서는 조정 R^2 값을 활용한다. 조정 R^2 값은 항상 R^2 값과 같거나 작게 산출되기 때문에 음수 값을 가질 수도 있다.

회귀 모형 및 (14-나)의 다른 후보 모형에 비해 더 높은 통계적 설명력을 보인다. <표 7>은 후보 모형의 R^2과 조정 R^2을 비교하여 제시한 것으로 다른 후보 모형의 조정 R^2 값은 [문단 길이 + 문장 길이 + 어절 길이]로 구성된 다중 회귀 모형에 비해 적게 산출된다. 이는 한국어 학습 교재의 텍스트 수준을 평가할 때 개별 가독성 지표만을 상정하는 것보다, 문단 길이, 문장 길이, 어절 길이를 함께 고려하는 것이 관측 자료의 특성을 보다 많이 반영할 수 있음을 의미한다. 물론 [문단 길이 + 문장 길이]와 [문단 길이 + 문장 길이 + 어절 길이] 모형의 조정 R^2 값은 차이가 크지 않으므로, 어절 길이의 효과가 적다고 할 수 있다. 그래서 문단 길이와 문장 길이만으로 구성된 한국어 학습 교재의 가독성 공식을 고려할 수도 있다. 이와 관련한 문제는 5.2 절에서 다시 논의하고, (15)의 공식을 수정하여 제시할 것이다.

5. 가독성 응용

4 절에서는 본 연구에서 관찰 대상으로 하는 가독성 지표, 즉, 문단 길이, 문장 길이, 어절 길이가 학습 단계에 따른 텍스트 수준과 상관성이 있음을 통계적으로 검증하였다. 이러한 통계적 상관성이 있다면, 한국어 학습 교재는 학습 단계 및 학습 단원에 따라 순차적인 텍스트 난이도의 변화가 대체로 반영되어 있다는 것을 의미한다. 즉, 본 논문의 연구 대상 자료인 한국어 학습 교재가 비록 명시적이고, 체계적으로 가독성 지표를 고려하여 설계된 것은 아닐지라도, 학습 수준을 고려하여 선별된 학습 단계별 텍스트는 일정 부분 가독성 지표의 난이도를 반영하고 있다고 할 수 있다. 이는 한편으로 학습 수준에 따른 텍스트 난이도를 측정할 수 있는 척도로서 가독성 지표를 활용할 수 있음을 나타낸다.[21] 5 절에서는 이러한 상관성을 토대로 다음의 두 가지 측면에서 한국어 학습 교재에 대한 통계적 평가 방법론을 논의한다. 첫째, 5.1 절에서는 학습 단계 및 학습 단원의 인접 유사성을 평가한다. 즉, 인접 학습 단계 및 인접 학습 단원 사이에 가독성 척도의 유사성이 고려되어 배치 순서에 따라 순차적인 가독성 지표의 변화가 정밀하게 고려되어 있는지를 평가한다. 둘째, 한국어 학습 교재에 대한 가독성 척도의 추세 구간을 평가한다. 비록 한국어 학습 교재가 순차적인 가독성 지표의 변화를 반영하고 있을지라도, 그 변화의 크기 및 방향이 모든 학습 단계에 걸쳐 동일하지 않을 수 있다.

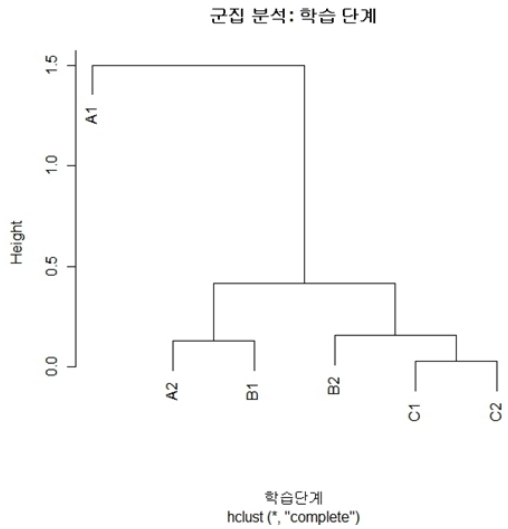
---

21) 물론 본 논문에서 다루는 가독성 지표만으로 텍스트 수준을 완전하게 측정할 수는 없다. 이 외에도 텍스트에 나타나는 어휘나 문법적 요소도 텍스트 수준 측정에 고려될 수 있을 것이다.

그래서 5.2 절에서는 학습 단계를 두 구간으로 구분하여 구간별 가독성 척도의 추세 및 그 추세 변환 특성을 평가한다.

5.1. 인접 유사성 평가

한국어 텍스트 교재가 순차적인 텍스트 난이도에 따라 적절하게 구성되어 있다면, 인접 학습 단계 및 인접 학습 단원의 텍스트는 가독성 지표의 관측치가 유사하거나, 또는 그 차이가 다른 텍스트에 비해 가장 적은 수치로 관찰될 것이다. 이러한 점에 착안하여 5.1 절에서는 학습 단계 및 학습 단원 배치의 적절성을 평가하기 위해 피어슨 적률 상관계수에 기반한 응집적 군집 분석을 활용한다.[22] 응집적 군집 분석은 가독성 측정치의 유사성에 따라 학습 단계 및 학습 단원을 동일 군집으로 분류하는데, 만약 교재가 순차적인 가독성 척도를 반영하고 있다면 인접 학습 단계 및 인접 학습 단원의 텍스트는 동일 군집으로, 그리고 순차적으로 군집화된 수형 구조도로 분석될 것이다.



<그림 4> 학습 단계별 텍스트의 군집 분석

<그림 4>는 문단 길이, 문장 길이, 어절 길이의 척도를 통합적으로 고려하여 학습 단계별 텍스트를 군집 분석한 결과이다. <그림 4>에서 유사한 특성의

22) R 통계 패키지에서 응집적 군집 분석은 hclust( ) 사용한다. 상관계수에 기반한 응집적 군집 분석의 자세한 사용법은 Baayen(2008)의 5.1.5 절 참조.

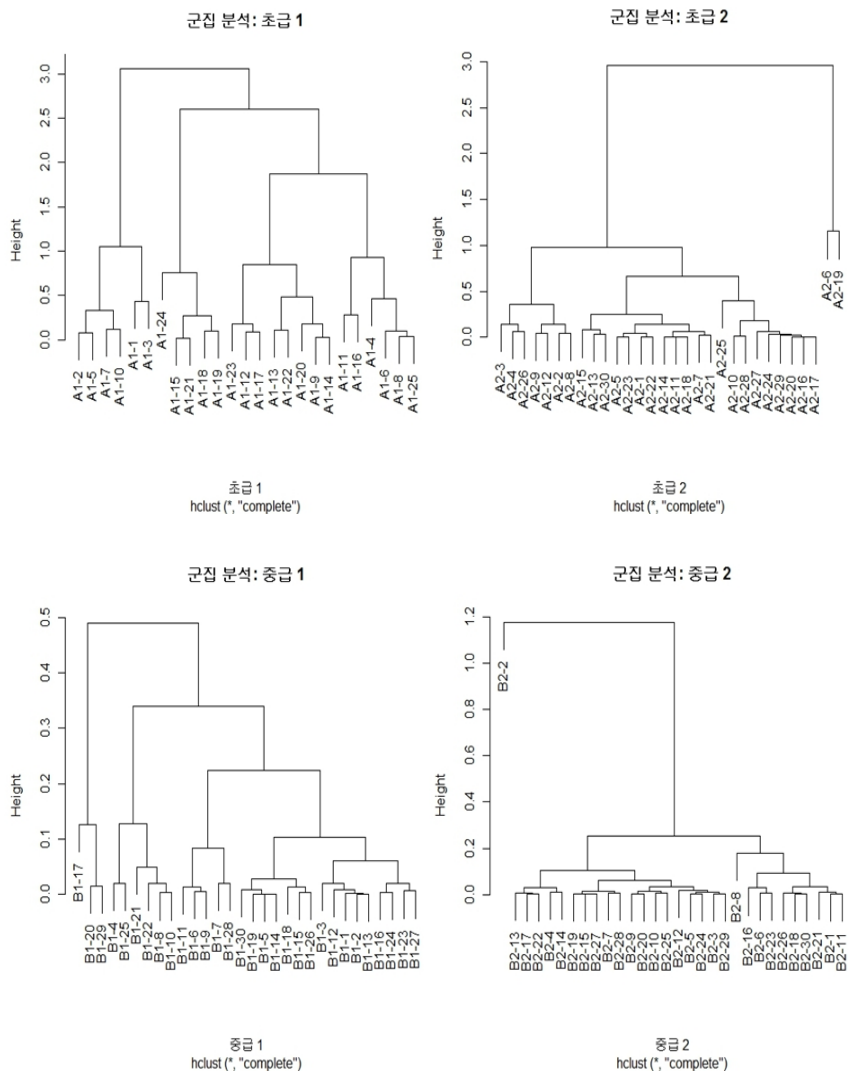
학습 단계는 수평선을 통해 동일 군집으로 분류되며, 수형 구조도의 아래에서부터 순차적으로 군집화하여 최종적으로 하나의 군집으로 분류된다. 이 수형 구조도에서 수직선의 거리는 비유사성의 크기를 의미한다. 초급-1 단계(A1)의 텍스트는 다른 학습 단계의 텍스트와 구분되는 위계를 보이고 있으며, 다른 학습 단계와 연결되는 수직선의 거리 또한 멀기 때문에 비유사성이 크다고 할 수 있다. 초급-1 단계 텍스트와 비유사성이 큰 학습 단계에서는 초급-2 단계(A2)와 중급-1 단계(B1)가, 고급-1 단계(C1)와 고급-2 단계(C2)가, 중급-2 단계(B2)와 고급 단계(C1, C2)가 유사성을 보이고 있다. 이에 따라 연구 대상 학습 교재는 인접 학습 단계 텍스트 사이의 유사성을 준수하고 있으며, 학습 단계에 따른 순차적인 가독성 척도의 차이를 반영하고 있다고 할 수 있다.

(16) 학습 단계의 군집 유형
- 초급1(A1)
- 초급2(A2), 중급1(B1)
- 중급2(B2), 고급1(C1), 고급(C2)

그러나 이를 크게 세 유형으로 구분한 (16)을 보면, 초급과 중급의 소분류 학습 단계는 동일 대분류 학습 단계보다 인접 대분류 학습 단계와 유사성이 크다. 다시 말해서, 초급-2 단계(A2)의 텍스트는 초급-1 단계(A1)보다 중급-1 단계(B1) 텍스트와, 중급-2 단계(B2) 텍스트는 중급-1 단계(B1)보다 고급 학습 단계의 텍스트와 유사성이 크다. 물론 본래의 교재 구성 취지가 (16)과 같다거나, 교재 구성에서 대분류 학습 단계의 엄격한 구분이 큰 의미를 갖는 것이 아니라면, 인접 학습 단계의 유사성은 분명하게 준수되어 있다고 할 수 있다. 그러나 교재 구성에서 대분류 학습 단계의 분류가 중요한 가치를 갖는 평가 요소라면 이에 대한 가독성 척도가 섬세하게 고려되어 있지 않은 것으로 판단된다.

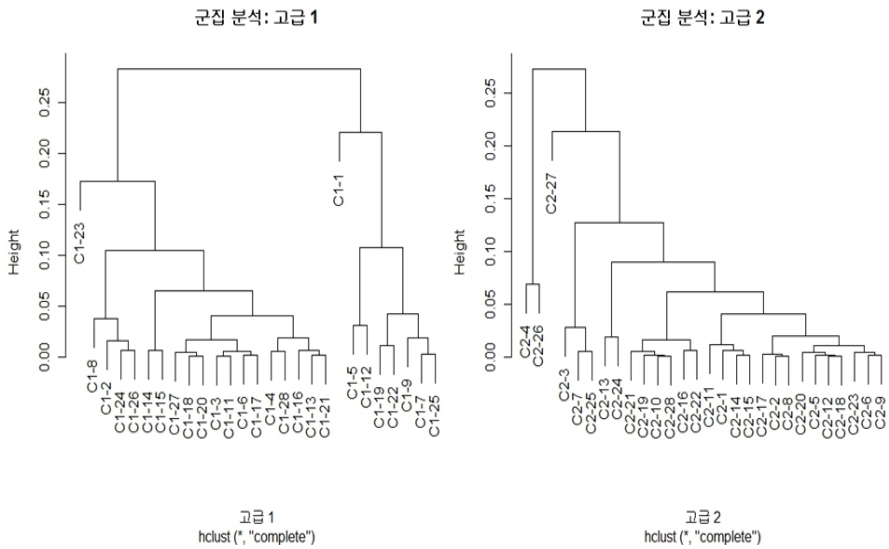
이러한 문제는 학습 단계별 텍스트의 단원 배치에서도 관찰된다. <그림 5>는 초급, 중급 단계에 따른 단원별 텍스트의 군집 분석 결과이다. 초급 및 중급 단계의 각 텍스트는 대체로 바로 인접한 단원보다는 비교적 인접한 단원들과 군집화되어 있다. 그러나 단원의 순서 차이가 큰 텍스트의 군집도 관찰된다. 초급-1 단계에서는 8 단원(A1-8)과 25 단원(A1-25)이, 초급-2 단계에서는 4 단원(A2-4)과 26 단원(A2-26), 13 단원(A2-13)과 30 단원(A2-30) 등이, 중급-1 단계에서는 4 단원(B1-4)과 25 단원(B1-25), 7 단원(B1-7)과

28 단원(B1-28) 등이, 중급-2 단계에서는 7 단원(B2-7)과 28 단원(B2-28), 3 단원(B2-3)과 28 단원(B2-29) 등이 이러한 예에 해당된다. 또한 수형 구조도에서 수직선 거리의 차이가 큰 단원도 일부 포함하고 있다. 특히, 초급-2 단계의 6 단원(A2-6)과 19 단원(A2-19), 중급-2 단계의 2 단원(B2-2)은 다른 단원들과 비교하여 가독성 척도의 비유사성이 크다. 이러한 측면에서 관찰 대상 교재의 단원 배치는 가독성 지표에 대한 고려가 완전하지 않다고 할 수 있다.



<그림 5> 초급 및 중급 단계 텍스트의 군집 분석

고급 단계 텍스트의 단원별 군집 분석 <그림 6> 또한 초급 및 중급 단계 텍스트의 분포와 크게 다르지 않다. 그런데 <그림 6>에서 문어 텍스트는 대화체 텍스트와 다른 군집으로 분석되어 있다. 3.1 절에서 언급한 것처럼 연구 대상 텍스트 총 171 단원 중 8 개 단원만이 문어 텍스트인데, 고급-1 단계에 6 개 단원 (17-가), 고급-2 단계에 2 개 단원 (17-나)가 분포한다. <그림 6>의 고급-1 단계 도식(좌측)에서 (17-가)의 문어 단원은 오른쪽 군집으로 분포되어 있다. 비록 연구 대상 단원 중 문어 텍스트의 비중이 적어 본 연구에서 대화체 텍스트와 문어 텍스트를 구분하여 관찰하지 않았지만, 이 분석 결과를 통해 문어 텍스트의 가독성 척도는 대화체 텍스트와 구분된다고 할 수 있다. 이에 비해 고급-2 단계의 문어 단원 (17-나)는 이러한 군집 경향이 분명하게 구분되지 않는다. 이는 고급-2 단계의 모든 대화체 텍스트에 대화 상황이 기술된 문어 텍스트가 혼용되어 있기 때문인 것으로 보인다.



<그림 6> 고급 단계 텍스트의 군집 분석

(17) 문어 단원
가. 고급-1 단계: 1, 5, 9, 12, 19, 22 단원
나. 고급-2 단계: 7, 26 단원

이상과 같이 한국어 학습 교재의 학습 단계 및 학습 단원은 순차적인 가독성 척도의 변화에 대한 정밀한 고려가 부족하다고 할 수 있다. 4 절에서 살펴본

것처럼 텍스트 수준과 가독성 지표의 상관성이 있으므로, 이에 따라 학습 단계 및 학습 단원을 구성할 때 가독성의 차이를 체계적으로 고려할 필요가 있어 보인다. 교재를 통한 학습 효과를 높이기 위해서는 학습자에게 적절한 수준의 텍스트를 순차적으로 제시하는 것은 중요한 평가 요소라 할 수 있다. 그러나 아직까지 이와 관련한 명시적 평가 기준 설정이 미흡한 상황이라 판단된다. 비록 이해영(2001)과 같이 주관적 판단에 의존하여 교재의 가독성을 평가한 일부 연구가 있으나, 주관적 판단에 의존하여 순차적인 가독성 척도를 체계적으로 평가하기는 어려워 보인다. 만약 한국어 학습 교재를 포함하여 다양한 학습 교재에 대한 평가에서 컴퓨터를 활용한 가독성 측정이 도입된다면, 군집 분석을 통해 교재 배치의 적절성을 객관적으로 평가할 수 있을 것이다.

5.2. 추세 구간 평가

5.1 절에서 군집 분석을 통해 인접 학습 단계와 인접 학습 단원의 유사성을 평가하였다. 그러나 이러한 접근은 텍스트의 배열 순서와 상관없이 가독성 척도의 포괄적인 유사성과 비유사성만을 평가할 뿐, 텍스트의 배열 순서에 따른 가독성 지표별 변화량 및 방향성은 제시하지 못한다. 물론 텍스트의 배열 순서에 따른 가독성 지표별 선형적 추세는 4.2 절의 상관 분석과 단순 선형 회귀 분석을 통해 파악할 수 있다. 그러나 이 통계 기법들은 관찰 대상 전 구간에서 전반적인 추세를 나타내는 하나의 선형 관계만을 도출하기 때문에 전반적인 추세에서 벗어나 있는 관측 자료의 분포적 특성을 불가피하게 누락시키기 마련이다. 실제로 4.2 절에서 제시했던 학습 단계와 가독성 지표에 대한 단순 회귀 분석의 R^2 값에 따르면, 회귀식으로 설명 가능한 관측 자료의 분포 비율은 문단 길이에서 47.27%, 문장 길이에서 44.62%, 어절 길이에서 16.59%로 해당 선형 모형으로 설명하기 어려운 관측 자료들의 분포 비율이 훨씬 높다.[23]

비록 관측 자료에 대한 특성을 선형 회귀식으로 완벽하게 설명하기는 어렵다고 하더라도, 관찰 구간을 세분화하여 구간별 회귀식을 산출한다면 적어도 두 가지 장점이 있다. 첫째, 관측 자료의 추세를 보다 자세하게 평가할 수 있을 것이다. 전 구간을 대상으로 추출된 회귀식은 전 구간에 걸쳐 동일 변동량의, 그리고 동일 방향성의 일관된 선형 관계만을 나타내는 반면, 구간별 회귀식은 구간에 따라 변동량 및 방향이 상이한 선형 관계를 나타낼 수 있기 때문에 관측 자료의 특성을 보다 자세하게 반영한다고 할 수 있다. 둘째, 회귀식으로

23) 그러나 4.2 절 (13)에서 제시한 R^2 값의 판별 기준(Cohen 1988)에 따르면 문단 길이 및 문장 길이의 선형 모형은 "높은 효과"로, 어절 길이는 "중간 효과"로 판별할 수 있다.

설명 가능한 관측 자료의 분포 비율, 즉, R^2 값을 향상시킬 수 있다. 구간별 회귀식은 관측 자료의 특성을 보다 자세하게 반영할 수 있으므로 설명 가능한 관측 자료의 분포 비율이 전 구간의 회귀식보다 높을 가능성이 있다. 구간별 회귀식의 이러한 특성을 학습 교재 평가에 적용한다면, 가독성 척도의 변화 추세를 달리하는 학습 구간을 탐색하는 용도로 활용할 수 있을 것이다.

5.2 절에서는 구간을 분할하여 구간별 회귀식을 도출하기 위해 Baayen(2008)에서 제시한 구분점을 포함한 선형 회귀 모형(regression with breakpoint)을 활용한다. Baayen(2008)은 관찰 구간을 두 구간으로 구분하고, 왼쪽 구간에 대한 선형 회귀식과 오른쪽 구간에 대한 선형 회귀식, 즉, 두 개의 구간별 선형 회귀식으로 구성되는 선형 회귀 모형 중에서 관측 자료와 가장 유사한 회귀 모형과 그 구간의 구분점을 산출하는 통계 기법을 제시하고 있다.

(18) 후보 선형 회귀 모형과 관측 자료의 편차 계산을 위한 R 코드[24)]

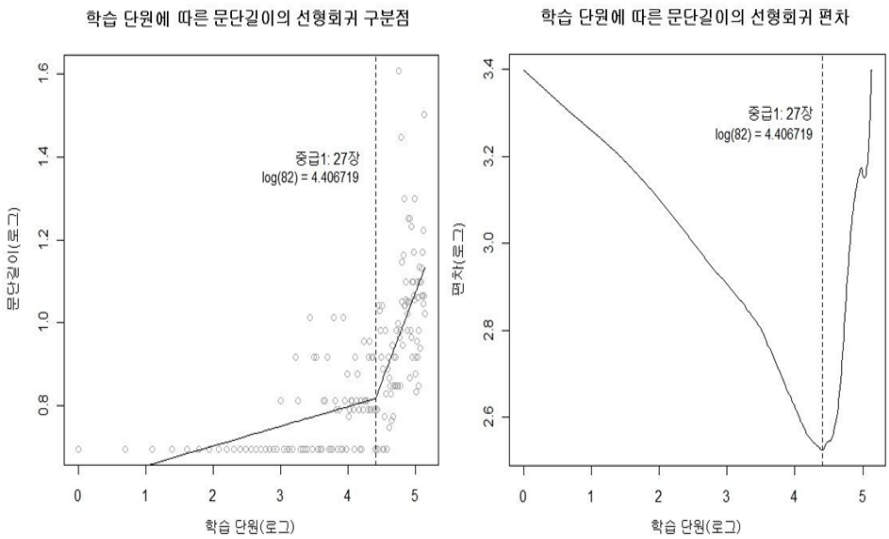
```
for (pos in 1:(nrow(DataVector)-1)) {
    breakpoint = log(pos)
    DataVector$ShiftedLogDistance = DataVector$LogDistance - breakpoint
    DataVector$PastBreakPoint = as.factor(DataVector$ShiftedLogDistance > 0)
    DataVector.both = lm(LogValue) ~ ShiftedLogDistance:PastBreakPoint,
                          data = DataVector)
    deviances[pos] = deviance(DataVector.both)
}
```

24) (18) R 코드의 변수와 설명
- pos: 하나의 구간을 갖는 후보 구분점 및 두 구간으로 분할 가능한 후보 구분점의 위치값을 나타내며, 전 구간의 크기만큼 처음부터 차례대로 구분점의 위치를 변환해 간다.
- DataVector: 종속 변수, 즉, 전 구간의 가독성 지표 값 목록
- breakpoint: 후보 구분점 위치의 로그값
- DataVector$ShiftedLogDistance: 후보 구분점을 기준으로 각 종속 변수 값의 좌우 위치
- DataVector$LogDistance: 전 구간에서 각 종속 변수 값의 순서를 로그로 변환한 값
- DataVector$PastBreakPoint: DataVector$ShiftedLogDistance의 값, 즉, 후보 구분점을 기준으로 한 각 종속 변수 값의 좌우 위치 값에 따라 True 또는 False 판정
- LogValue : 독립 변수, 즉, 학습 단원 순서의 로그 변환 값
- DataVector.both: 후보 구분점을 기준으로 좌우 두 개의 선형 회귀식으로 구성된 후보 선형 회귀 모형
- deviances: 후보 선형회귀 모형과 관측 자료 사이의 분산 값 목록으로, 이 분산 값 목록 중에서 최소값을 나타내는 후보 구분점의 선형 회귀 모형이 최종적으로 최적 모형으로 선정된다.

(19) (18) 코드 내용 요약

첫째, 후보 구분점을 전 구간에 걸쳐 순차적으로 상정
둘째, 각 후보 구분점을 기준으로 좌우 두 구간에 대한 선형 회귀식으로 구성된 후보 선형 회귀 모형을 산출
셋째, 후보 선형 회귀 모형과 관측 자료의 편차를 편차 목록에 저장

(18)은 관측치와 가장 유사한 선형 회귀 모형을 산출하기 위해, 각 후보 구분점을 기준으로 하여 생성된 후보 회귀 모형과 관측 자료의 편차를 계산하는 R 통계 패키지의 코드이다. (19)는 (18)의 코드 내용을 간략하게 정리한 것이다. (18)의 코드를 처리한 후 편차 목록(변수명: deviances) 중에서 최소의 편차값을 나타내는 후보 구분점과 후보 선형 회귀 모형을 최적의 모형으로 추출한다. 여기서 최소값의 편차를 나타내는 후보 구분점을 선택하는 이유는 이 구분점을 기준으로 구간을 분리하여 산출한 선형 회귀 모형이 후보 모형 중에서 관측 자료와 가장 근접한 것이기 때문이다.[25]



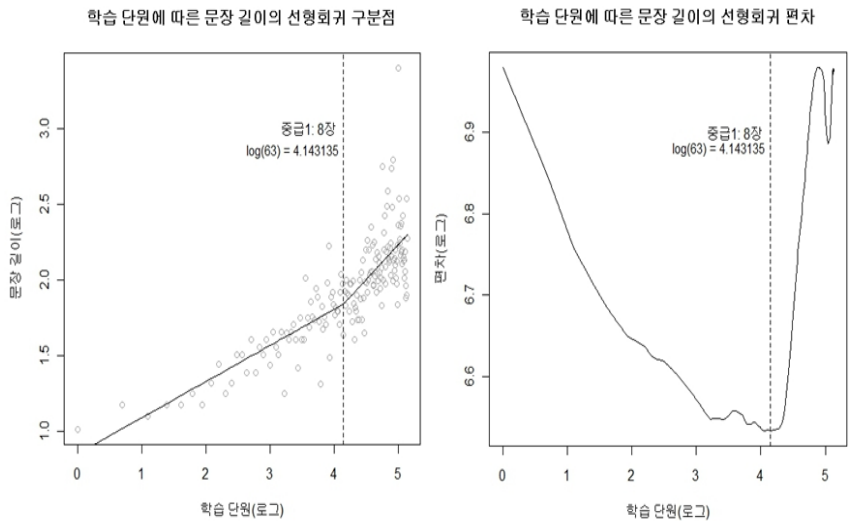
<그림 7> 문단 길이: 구분점 선형 회귀 모형(좌)과 편차(우)

<그림 7>은 (18)의 처리를 통해 최적의 구분점 및 선형 회귀 모형으로 산출된 문단 길이의 구간별 선형 회귀 모형(좌)과 편차(우)이다. <그림-7>의

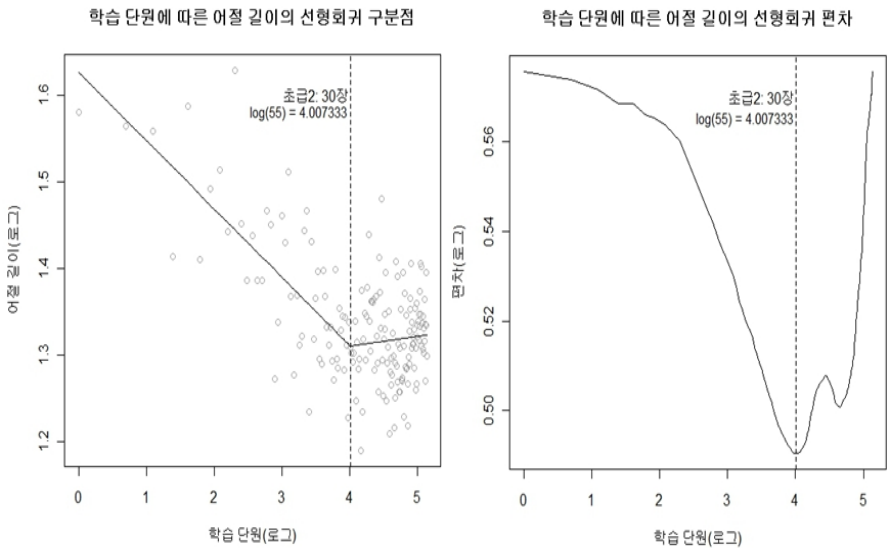
25) 구분점을 포함한 회귀 모형에 대한 자세한 내용은 Baayen(2008)의 6.4 절 참조.

왼쪽 도식에서 문단 길이의 구분점은 중급-1 단계의 27 단원에 위치하며, 이 구분점을 기준으로 오른쪽 구간의 선형 모형은 왼쪽 구간의 선형 모형에 비해 가파른 상승세를 보이고 있다. 이는 학습 단원에 따라 전체적으로 문단 길이의 증가 추세를 보이기는 하지만, 특히, 중급-2 단계부터 고급-2 단계까지의 구간에서 문단 길이가 가파르게 증가하고 있음을 나타낸다. 또한 <그림 7>의 오른쪽 도식에서 이 학습 단원을 구분점을 기준으로 하여 구간을 분할한 선형 회귀 모형과 관측 자료 사이의 편차는 다른 학습 단원을 구분점으로 하여 산출된 편차와 비교하여 가장 적은 값을 나타낸다. 이는 중급-1 단계의 27 단원을 기준으로 분할된 두 구간의 선형 회귀식이 문단 길이의 관측 자료를 가장 잘 반영함을 의미한다.

<그림 8>과 <그림 9>는 각각 문장 길이와 어절 길이에 대한 구간별 선형 회귀 모형(좌)과 편차(우)이다. 문장 길이는 중급-1 단계의 8 단원을, 어절 길이는 초급-2 단계의 30 단원을 구분점으로 한다. 문장 길이에서도 문단 길이와 마찬가지로 구분점 오른쪽 구간의 선형 모형이 왼쪽 구간의 모형보다 우세한 상승세를 보이고 있어 대체로 중급-1 단계 초반부터 문장 길이의 상승률이 높다고 할 수 있다. 반면, 어절 길이의 경우 구분점 왼쪽 구간의 선형 모형, 즉, 초급-1 단계와 초급-2 단계는 하향 추세를 보이지만, 오른쪽 구간의 선형 모형, 즉, 중급-1 단계부터 고급-2 단계까지는 상승이나 하강의 추세가 명확하게 관찰되지 않는다.



<그림 8> 문장 길이: 구분점 선형 회귀 모형(좌)과 편차(우)



<그림 9> 어절 길이: 구분점 선형 회귀 모형(좌)과 편차(우)

(20) 가독성 지표별 구분점

- 문단 길이: 중급-1 단계 27 단원(총 30 단원 중)
- 문장 길이: 중급-1 단계 8 단원(총 30 단원 중)
- 어절 길이: 초급-2 단계 30 단원(총 30 단원 중)

초급-2 단계, 중급-1 단계의 총 30 단원 구성을 감안하여 가독성 지표별 구분점 (20)을 보면, 연구 대상 자료의 구분점은 대체로 중급-1 단계를 기점으로 추세가 변환되는 것으로 파악된다. 그래서 중급-2 단계, 고급-1 단계, 고급-2 단계는 문단 길이 및 문장 길이에서 이전 구간에 비해 가파른 상승세를 보이며, 어절 길이의 선형적 변동성은 거의 관찰되지 않는 유형으로 분류될 수 있다. 이에 비해 초급-1 단계 및 초급-2 단계는 문단 길이 및 문장 길이의 비교적 완만한 상승과 어절 길이의 하향 추세를 공통적 특징으로 한다.

이렇게 학습 단계가 가독성 척도에 의해 두 구간으로 분리될 수 있다면, 변환 기점이 되는 중급-1 단계 학습자는 가독성 척도의 급격한 변화에 따른 학습적 어려움을 체감할 수도 있다. 특히, 문단 길이, 문장 길이가 이전 구간에 비해 크게 상승하므로, 가독성 척도의 변화에 대해 적응이 어려울 수도 있다. 그러므로 교재 구성에서 이러한 특성을 고려하여 중급-1 단계 학습 텍스트의 가독성 척도를 조정하거나, 보충 읽기 자료 등을 통해 중급-1 단계 학습자의 오른쪽 구간에 대한 텍스트 적응력을 향상시킬 필요가 있어 보인다.

구분점을 포함한 선형 모형의 이러한 특성은 5.1 절 (16)에서 제시된 학습 단계의 군집 유형과도 관련된다. 5.1 절 (16)의 학습 단계의 군집 유형에서 중급-2 단계, 고급-1 단계, 고급-2 단계가 동일 군집으로 분류되는 점도 이러한 특성에 기인한 것으로 판단된다. 즉, 이들 학습 단계의 텍스트는 문단 길이, 문장 길이, 어절 길이에서 모두 동일 선형 관계 구간 내에 위치하여 가독성 척도의 특성이 유사하다고 하겠다. 반면, 중급-1 단계의 텍스트는 문단 길이와 문장 길이와 관련하여 두 구간에 속하는 단원들이 모두 포함되어 있으므로 위 유형의 학습 단계들과 구분된다. 그래서 학습 단계의 군집 유형 (16)에서 중급-1 단계의 텍스트는 동일 대분류 학습 단계인 중급-2 단계가 아닌, 인접 대분류 학습 단계인 초급-2 단계와 동일 군집으로 분석되는 것으로 보인다.

<표 8> 구분점 선형 회귀 모형의 분석 결과

	계수	t-value	p-value	F-value	p-value	조정 R^2	단순회귀 조정 R^2
절편	0.81619	49.916	< 2e-16				
좌 모형	0.0475	3.494	0.000609				
우 모형	0.43266	10.171	< 2e-16				
문단길이				99.01	< 2.2e-16	0.537	0.4547
절편	1.84061	65.492	<2e-16				
좌 모형	0.23884	9.689	<2e-16				
우 모형	0.45979	9.338	<2e-16				
문장길이				183.5	< 2.2e-16	0.6835	0.4315
절편	1.310591	164.528	<2e-16				
좌 모형	-0.0788	-10.951	<2e-16				
우 모형	0.011338	0.956	0.341				
어절길이				74.44	< 2.2e-16	0.465	0.1733

한편, <표 8>은 가독성 지표별 구분점을 포함한 선형 회귀 모형의 분석 결과이다. 각 선형 모형은 구분점을 기준으로 분할된 왼쪽 구간 모형(좌 모형)과 오른쪽 구간 모형(우 모형)으로 구성되어 있으며, 어절 길이의 우 모형 p-value를 제외한[26] 모든 가독성 지표의 절편, 좌 모형, 우 모형, F 통계량의 p-value가

26) 선형 회귀 분석에서 절편 및 회귀계수에 대한 t-value의 p-value가 유의 수준보다 높다면, 해당 계수의 값이 0임을 의미한다. 특히, 회귀계수가 0이라면 기울기가 0인 직선을 나타내므로

유의수준 0.001보다 작다. 이는 어절 길이의 오른쪽 구간 모형(우 모형)은 선형 관계를 갖지 않음을, 다시 말해서 선형의 변화량 및 방향성을 나타내지 않음을 나타내며, 이를 제외한 가독성 지표별 왼쪽 구간 모형(좌 모형)과 오른쪽 구간 모형(우 모형)은 모두 선형적 관계가 통계적으로 유의미하며, 가독성 지표별 선형 모형 또한 통계적으로 유의미함을 나타낸다.

<표 8>에서 다른 지표와 달리 어절 길이는 오른쪽 구간의 선형 모형에서 통계적으로 유의미한 선형성을 보이지 않는다. 즉, 중급-1 단계부터 고급-2 단계까지는 어절 길이의 선형 관계가 통계적으로 관찰되지 않는다고 할 수 있다. 이러한 특성은 4.2 절에서 제시했던 어절 길이의 상관 분석 및 단순 선형 회귀 분석과 차이가 있다. 구간을 분할하지 않고, 전 구간을 대상으로 어절 길이와 텍스트 수준의 상관성을 상관 분석 및 단순 선형 회귀 분석을 통해 측정했던 4.2 절에서는 어절 길이를 통계적으로 유의미한 가독성 지표로 제시하였다. 그러나 중급 및 고급 단계에서 어절 길이의 선형 관계가 관찰되지 않으므로, 전 구간을 대상으로 한 가독성 지표로 어절 길이를 상정하는 것은 문제가 있어 보인다. 또한 한국어 학습 교재의 가독성 공식을 도출하기 위해 4.2 절 <표-7>에서 제시했던 후보 선형 모형의 조정 R^2 값에서도 [문단 길이 + 문장 길이 + 어절 길이]와 [문단 길이 + 문장 길이]의 조정 R^2 값은 각각 0.6398과 0.6113으로 근소한 차이를 보이므로, 일부 구간에서 선형성을 보이지 않는 어절 길이를 다중 회귀 모형에서도 고려할 필요가 없어 보인다.

(21) 한국어 학습 교재의 수정 가독성 공식[27]

- $Y = -0.3897 + 1.5995X_1 + 0.2471X_2$
- X_1 : 문단 길이, X_2 : 문장 길이

따라서 4.2 절에서 [문단 길이 + 문장 길이 + 어절 길이]로 구성된 가독성 공식을 수정하여 어절 길이를 제외한 [문단 길이 + 문장 길이]로 구성된 다중

이는 선형 관계를 갖지 않음을 의미한다.

27) 한국어 학습 교재의 수정 가독성 공식 (21)에 대한 다중 회귀 분석 결과는 다음과 같다.

모형	계수	t-value	p-value	F-value	p-value	조정 R^2
절편	-0.3897	-1.505	0.134			
문단 길이	1.5995	8.590	5.89e-15			
문장 길이	0.2471	7.893	3.72e-13			
				133.9	< 2.2e-16	0.6113

회귀 모형을 한국어 학습 교재의 수정 가독성 공식 (21)로 제안한다.

그러나 비록 어절 길이는 한국어 학습 교재의 전체 학습 단계를 대상으로 한 가독성 지표에서 제외되지만, 구분점 선형 모형의 왼쪽 구간인 초급 단계에서는 어절 길이가 통계적으로 유의미한 선형 관계를 나타내므로, 초급 단계에서는 여전히 유효한 가독성 지표라 할 수 있다. 이러한 특성은 초등학교 교과서에서도 유사하게 관찰되고 있다. 최인숙(2005)는 어절 길이가 초등학교 텍스트를 포함하여 초중고 교과서를 대상으로 한 가독성 측정에서 텍스트 수준과 통계적 상관성을 보이는 반면, 초등학교 텍스트를 제외한 중고등학교 교과서를 대상으로 한 측정에서 유의하지 않다고 한다. 즉, 어절 길이가 초등학교 교과서에서 통계적으로 유의미한 상관성을 보인다고 간접적으로 판단할 수 있다. 이와 같이 어절 길이는 한국어 학습 교재의 초급 단계 및 초등학교 수준의 텍스트를 포함한 초중고 교과서에서 유효한 가독성 지표인 것으로 보아, 초보적인 학습 텍스트에서만 통계적으로 유효한 가독성 지표일 가능성이 있다. 이러한 측면에서 학습 단계별로 적용 가능한 가독성 지표를 구분한다면, 초급 단계에서는 다른 학습 단계와 달리 문단 길이 및 문장 길이와 더불어 어절 길이도 텍스트 수준을 구분하는 가독성 지표로 상정할 필요가 있다.

이 밖에도 <표 8>에서 구분점을 포함한 선형 회귀 모형의 조정 R^2 값은 4.2 절에서 제시한 단순 선형 모형의 조정 R^2 값에 비해 높게 산출된다.[28] 단순 선형 모형의 조정 R^2 값과 구분점 선형 회귀 모형의 조정 R^2 값은 각각 문단 길이에서 0.4547과 0.537로, 문장 길이에서 0.4315와 0.6835로, 어절 길이에서 0.1733과 0.465로 분석된다. 이는 구분점 선형 회귀 모형이 단순 선형 회귀 모형에 비해 관측 자료의 특성을 잘 반영하고 있으며, 향상된 통계적 설명력을 보이고 있음을 나타낸다.

이상과 같이 구분점을 포함한 선형 모형은 관찰 구간을 분할하여 두 가지 선형 관계를 나타냄으로써, 첫째, 단순 선형 모형에 비해 관측 자료에 대한 통계적 설명력을 향상시키고, 다중 선형 모형에서 파악하기 어려운 구간별 비선형성을 제시하여 관측 자료의 보다 정확한 선형 관계를 반영하고 있으며, 둘째, 5.1 절의 군집 분석을 통해 파악하기 어려웠던 가독성 지표별 유사성 및 비유사성 학습 단계에 대한 선형 관계적 특성을 제시할 수 있다. 이러한 특성을 교재 평가에 활용하면 보완이 필요한 가독성 지표 및 학습 단계 탐색, 그리고 보완 방향 제시 등이 가능할 것으로 기대된다.

---

28) 물론 구분점을 포함한 선형 모형은 학습 단원을, 4.2 절의 단순 선형 모형은 학습 단계를 변수로 하므로 차이가 있다. 그러나 학습 단계를 독립 변수 또는 의존 변수로 하는 단순 선형 회귀 모형이나, 학습 단원을 의존 변수로 하는 단순 선형 회귀 모형이나 조정 R^2 값은 동일하므로, 본 논문에서는 이에 대한 별도의 분석 결과를 제시하지 않겠다.

6. 결론

지금까지 한국어 학습 교재의 텍스트 수준과 문단 길이, 문장 길이, 어절 길이 사이의 통계적 상관성을 검증하고, 이 가독성 지표를 활용한 통계적 교재 평가 방법론에 대해 논의하였다. 먼저 상관 분석, 선형 회귀 분석, 구분점을 포함한 선형 회귀 모형을 통해 문장 길이와 어절 길이는 텍스트 수준에 따라 지속적으로 길어지는 경향을, 어절 길이는 초급 단계에서 점차 짧아지지만, 중급 단계에 접어들어서 선형성을 파악할 수 없는 경향을 관찰하였다. 어절 길이의 이러한 특성으로 인해 문단 길이 및 문장 길이만을 전체 학습 단계에 대한 유효한 가독성 지표로 상정해야 하지만, 초급 단계만을 대상으로 한 텍스트 수준 평가에서는 어절 길이를 유효한 가독성 지표로 활용할 수 있다.

또한 가독성 지표에 기반하여 군집 분석 및 Baayen(2008)의 구분점을 포함한 선형 회귀 모형을 통해 학습 단계 및 학습 단원에 대한 배치의 적절성을 평가하였고, 가독성 지표별 선형 관계 및 그 특성을 두 구간으로 구분하여 평가하였다. 특히, 구분점을 포함한 선형 모형을 통해 관찰 자료에 대한 통계적 설명력이 단순 선형 회귀 분석에 비해 향상되는 특성을 살펴보았고, 상관 분석, 단순 선형 회귀 분석, 다중 선형 회귀 분석에서 관찰하기 어려운 구간별 선형적 추세 및 비선형성을 용이하게 포착할 수 있었다.

아직까지 한국어 텍스트에 대한 가독성 측정 및 활용 연구가 활발하게 진행되지 못한 측면이 있지만, 이러한 통계적 가독성 측정 및 활용은 텍스트 수준 평가에 크게 기여할 것으로 기대된다. 무엇보다 가독성 처리는 컴퓨터의 이용을 기반하므로 방대한 분량의 텍스트를 빠르면서도 용이하게 처리할 수 있으며, 텍스트를 객관적으로 일관되게 평가할 수 있는 이점을 가지고 있다. 특히, 한국어 학습 교재를 비롯하여 다양한 분야의 학습 교재는 학습자 및 학습 단계에 적절한 수준의 텍스트로 구성되는 것이 바람직하므로, 교재 평가에서 가독성 측정 및 활용을 적극적으로 고려할 필요가 있어 보인다.

그러나 가독성의 활용도를 높이기 위해서 다음과 같은 연구가 향후 추진되어야 할 것이다. 첫째, 한국어 특성상 어절은 어휘와 문법 형태소가 결합되어 구성되므로 어절 길이에 어휘 난이도 및 형태 문법적 난이도가 반영되어 있을 가능성이 있다. 비록 본 연구에서 제한적인 가독성 지표로 관찰되고 있지만, 어절 길이에 대한 다각적인 논의를 통해 본질적인 특성을 규명할 필요가 있어 보인다. 둘째, 어휘 수준을 비롯하여 텍스트 수준 평가와 관련된 다양한 언어적 변인에 대한 검토 및 통계적 검증이 필요하다. 물론 가독성

측정은 컴퓨터를 이용해야 하므로, 연구 대상 언어적 변인은 컴퓨터를 이용하여 자동 분석이 가능한 것들이어야 한다. 셋째, 일반화된 가독성 척도를 구성하기 위해 연구 대상 범위를 확장하여 외국인 학습자 및 모국어 화자를 위한 표준화된 가독성 척도 및 공식이 개발되어야 할 것이다. 이를 위해서 다양한 한국어 학습 교재를 대상으로 하여 한국어 학습 교재에 공통적으로 적용할 수 있는 가독성 공식 연구가 수행되어야 하며, 아울러 모국어 화자의 가독성 척도에 대한 비교 연구도 수반되어야 할 것이다. 넷째, 텍스트 수준 측정 및 교재 평가에 활용할 수 있는 도구 개발 연구도 향후 진행되어야 할 것이다.

참고문헌

국립국어연구원. 2003. 한국어 학습용 어휘 선정 결과 보고서, 국립국어연구원 보고서.

심재홍. 1991. “글의 이독성에 영향을 미치는 요인과 이독성 측정의 모형화에 관한 연구.” 서울대학교 석사학위 논문.

이해영. 2001. “한국어 교재의 언어 활동 영역 분석.” 한국어교육 12-2, 469-490.

전정재. 2001. 독서의 이해, 서울: 한국방송출판(주).

진재관 외. 2008. “교과용 도서 평가 연구(II): 평가모형 개발 및 평가기준 설정.” 한국교육과정평가원 연구보고서.

진재관 외. 2009. “교과용 도서 평가 연구(III): 평가 도구 개발 및 적용.” 한국교육과정평가원 연구보고서.

최인숙. 2005. “텍스트의 언어적 난이도 측정 공식 비교 연구: 초중고 교과서를 중심으로.” 정보관리학회지 22-4, 173-195.

최재완. 1995. “신문 경제기사의 독이성에 관한 연구.” 경희대학교 박사학위 논문.

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

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*, Cambridge University Press.

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, 37-54.

Flesch, R. 1946. *The art of plain talk*. New York: Harpers.

Fry, E. 1977. “Fry's readability graph: Clarifications, validity and extension to level 17.” *Journal of Reading* 21-3, 242-252.

Gunning, R. 1952. *The Technique of Clear Writing,* New York: McGraw-Hill.

Heilman, M., K. C. Thompson, J. Callan, & M. Eskenazi. 2007. “Combining

lexical and grammatical features to improve readability measures for first and second language texts." *Proceedings of the Human Language Technology Conference,* 460-467.

Kincaid, J. P., R. P. Fishburne, R. L. Rogers, & B. S. Chissom. 1975. "Derivation of new readability formulas (Automated readability index, Fog count, and Flesch Reading Ease formula) for Navy enlisted personnel."Research Branch Report 8-75, Chief of Naval Technical Training: Naval Air Station Memphis.

Mclaughlin, H. 1969. "SMOG grading: A new readability formula." *Journal of Reading* 12-8, 639-646.

Si, L. & Callan, J. 2001. "A statistical model for scientific readability." *Proceedings of the Tenth International Conference on Information and Knowledge Management,* 574-576.

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

Stenner, A. J., I. Horabin, D. R. Smith, & R. Smith. 1988. *The Lexile Framework.* Durham, NC: Metametrics.

Sticht, T. G. 1973. "Research towards the design, development and evaluation of a job-functional literacy training program for the US Army." *Literacy Discussion* 4-3.

Warming, E. O. & Barber, E. C. 1980. *Touchstones for Textbook Selection*, Phi Delta Kappan.

Zenger, W. E. & Zenger, S. K. 1976. *Handbooks for Evaluating and Selecting Textbooks*, Belmont, California: Fearon Publishers.

▪ 홍정하 (HONG, Jungha)
소속: 고려대학교 언어정보연구소
(Research Institute for Language & Information, Korea University)
전자우편: kleist@korea.ac.kr

▪ 최재웅 (CHOE, Jae-Woong)
소속: 고려대학교 문과대학 언어학과
(Department of Linguistics, College of Liberal Arts, Korea University)
전자우편: jchoe@korea.ac.kr

▪ 유석훈 (YOU, Seok-Hoon)
소속: 고려대학교 문과대학 언어학과
(Department of Linguistics, College of Liberal Arts, Korea University)
전자우편: syou@korea.ac.kr

접수: 2011.02.08
게재결정: 2011.03.21