http://dx.doi.org/10.20880/kler.2017.52.3.135

토픽 모델링에 따른 고등학생 논설문의 응결성과 응집성의 상관분석

이슬기 한국교원대학교 국어교육과

I. 머리말

이 연구는 토픽 모델링을 사용하여 산출한 학생 글의 응결성(cohesion)[1] 지수와 응집성(coherence) 점수와의 상관관계를 분석하고, 응결성 측정 방법에 대한 타당성을 고찰하는 데 목적이 있다. 이를 위해 토픽 모델링(topic modeling)이라는 방법을 사용하여 학생 글 분석에 필요한 '응결성 지수'라는 개념을 제안하고, 이 논문에서 제안하는 응결성 지수가 응집성, 즉 글의 질적 특성을 얼마나 잘 짚어 낼 수 있는지에 대해 검증하려고 한다.

응집성은 일반적으로 응결성과 밀접한 관련을 맺으며 함께 논의되곤 했다. 응결성이 텍스트 기저의 의미 내용을 드러내기 위해 구성 성분들이 표층 구조에서 맺는 통사적 관계라면, 응집성은 텍스트 구성 요소들 간의 관계에서 비롯되는 긴밀한 의미적 관련성을 뜻한다. 응결성과 응집성은 텍스트다

1) 'coherence'와 'cohesion'은 연구자마다 '결속성, 결속 기제, 응집성, 일관성, 응결성' 등 다양한 용어로 번역하여 사용하였다. 국어과 교육과정에서는 '통일성'과 '응집성'으로 번역하고 있다. 본고에서는 보다 엄밀한 구분을 위해 한국텍스트 언어학회(2004)의 용어를 따르기로 한다.

움을 드러내는 언어적 조건과 의미적 조건인바 그간 국어교육에서도 그 중요성이 강조되어 왔고, 그로 인해 많은 연구자들에 의해 다방면의 연구가 축적되었다. 하지만 이 둘의 관계에 대해 범주화하여 글 수준에서 실증적으로 분석한 선행 연구는 찾아보기 어려웠다. 또한, 그간의 응결성과 응집성의 연구는 사람의 '읽는' 행위에서 시작했기 때문에 연구 결과에 주관적 영향력을 배제하기 어려웠다.

그러나 최근 기술의 약진에 힘입어 프로그램을 활용한 다양한 텍스트 분석 방법들을 사용할 수 있게 되었다. 텍스트 분석의 가장 큰 장점은 글에 대한 객관적 정보 제공이 가능하다는 것이다. 문장, 문단, 글의 수준에 따라 응결성을 각각 국지적 응결성, 포괄적 응결성, 글 전체의 응결성으로 구분할 때, 문장 수준에서는 문장 간 단어의 중복을, 문단 수준에서는 단락 간 의미의 유사성을, 글 전체 수준에서는 글 전반에 사용된 단어의 양상을 통해 응결성 측정이 가능한데, 이러한 유사 정도는 토픽 모델링을 통해 수치화할 수 있다. 이 연구에서는 이렇게 산출한 지수를 각각 국지적 응결성, 포괄적 응결성, 글 전체 응결성 지수로 설정하고, 교사 채점자들이 평가한 응집성 점수와의 상관관계를 분석하고자 한다. 응결성과 응집성의 분석에서 유의한 상관관계를 확인할 수 있다면, 토픽 모델링을 통한 자동 산출 방법이 학생 글에 대해 정량화된 정보를 제공하는 도구로써의 가능성을 타진해 볼 수 있을 것이다.

이 연구에서는 응결 장치(cohesive tie)를 토픽 모델링으로 산출한 '토픽의 유사성'에 한정하여 측정하였다. 따라서 지시나 대용 같은 다양한 응결 장치를 포함한 분석에까지 나아가지는 못했다는 점, 한 편의 글이 지닌 응결성을 단어로 분절하여 측정하고자 했다는 점의 한계를 갖는다.

그럼에도 불구하고 이 연구는 응결성을 범주로 구분하여 응집성과의 상관관계를 나타내 보고자 했다는 점에서 연구의 의의를 찾을 수 있다. 뿐만 아니라 방법론적인 면에서도, 그간 국어교육에서 사용하지 않았던 토픽 모델링을 활용하여 문장, 문단, 글 전체의 유사도를 산출했다는 점, 학생 글 분석에 준 지도 학습 방법을 이용한 기계 학습 방식을 새롭게 도입했다는 점에

서 유의미한 시도라 할만하다.

II. 이론적 배경

1. 텍스트 응집성과 응결성

텍스트성(textuality)은 통화의 과정에서 인간 정신, 언어, 현실적 국면이 텍스트에 개입되어 발현되는 것으로 상황성(situationality), 정보성(informativity), 의도성(intentionality), 수용성(acceptability), 상호 텍스트성(intertextuality), 응집성(coherence), 응결성(cohesion)을 기준으로 한, 텍스트가 갖추어야 할 7가지 특성을 말한다(Beaugrande & Dressler, 2008). 즉, 텍스트를 매개로 의사소통을 가능하게 하는 가장 근본적인 구성 원리가 텍스트성인 것이다.

이러한 텍스트성의 원리 중, 응결성은 텍스트 표층에서의 문법적 연관성을, 응집성은 텍스트 기저의 의미적 연관성을 나타내는 언어적 차원의 텍스트적 요인이다. 대체로 그간의 연구들은 응집성에 비해 응결성을 다룬 연구가 상대적으로 많았는데 응결성은 지시, 대용, 생략, 접속, 어휘적 응결성 같은 응결 장치를 통해 표층에 드러나지만, 응집성은 심층적 차원의 개념들과 관련되는바, 가시화되기 어렵기 때문이었다(박영민, 2004; 박진용, 2003).

연구자들에 따라 응집성과 응결성을 구분하지 않은 채 사용하기도 하고 한 개념을 다른 개념 안에 포함된 관계로 상정하기도 한다(Brinker, 1988; Halliday & Hasan, 1976). 국내에서도 응집성과 응결성의 구분에 대해 다양한 이견이 지속되어 왔다. 응집성과 응결성을 별개의 개념으로 서술하기도 하고(김정숙, 1996; 박영순, 2008) 응결성을 통해 응집성의 양상을 살펴본 연구도 다수 있었다(서승아, 2008; 이신형, 2012; 정다운, 2008). 이러한 연구들

은 대부분 총체적 평가 방식으로 응결성과 응집성의 상관관계에 대해 논했다. 최근 응결 장치를 대용, 접속, 상술로 나누어 응집성과의 상관도를 분석한 연구가 있었지만, 여기에서도 응결성의 범주를 구분하여 측정하지는 않았다(김민영, 2014; 박혜진·이미혜, 2017).

응결성은 국지적 응결성(local cohension), 포괄적 응결성(global cohension), 글 전체 응결성(overall text cohension)으로 세분할 수 있다(Crossley & McNamara, 2012). 국지적 응결성은 미시적 차원에서 문장과 문장 사이의 결속을 나타내며 문장들 간의 중복되는 단어나 연결어 사용을 통해 명시적으로 나타낸다. 이에 반해, 포괄적 응결성은 단락 수준에서 이루어지며 텍스트 단락 간 의미적 유사성에 의해 알 수 있다. 글 전체 응결성은 텍스트의 특정 부분에 대한 비교가 아닌, 텍스트 전반에 사용된 어휘의 다양성 같이 텍스트 전체의 특징으로 드러난다.

한편, 응집성은 독자가 응결성의 단서를 선행 지식이나 읽기 전략 같은 비언어적 요인과 결합하여 이해를 도출해 내는 것을 말한다. 따라서 텍스트 내부에 의미적으로 연결된 망으로서 존재하기 때문에 대부분 총체적 평가로 파악된다.

2. 토픽 모델링

텍스트 마이닝(text mining)은 비정형 데이터인 언어로 작성된 문서에서 유의미한 정보를 추출해 내는 새로운 텍스트 분석 방법을 말한다. 텍스트 마이닝에서는 단어를 기본 분석 단위로 하여 문서 내 출현 빈도, 단어들 간의 문서 내 동시 출현 확률 등을 계산하여 정보를 파악한다.[2] 텍스트 마이닝에는 여러 기법이 있는데, 토픽 모델링은 텍스트에 쓰인 단어들로부터 의미

2) 공기(co-occurrence)어 분석이라고도 불리는 동시출현 단어 분석(co-word analysis)은 분석 단위 안에 동시에 출현하는 횟수를 통해 단어 간 연관성이나 문서의 특징을 파악한다.

를 추출해서 텍스트 데이터에 대한 깊이 있는 분석을 가능케 한다는 점에서 최근 주목받고 있는 방법이다.

토픽 모델링은 여러 개의 문서가 모여 있는 문서 집합에서 특정 화제를 기준으로 토픽들을 군집화하기 위한 통계 모형으로, 문서에서 중요한 토픽을 확률적으로 계산해 내는 분석 방법이다. 토픽 모델링에서 화제는 각 화제를 구성하는 핵심어 목록에 의해 결정되고, 핵심어 목록이 빈번하게 공통적으로 분포하는 텍스트는 동일 화제의 텍스트로 분류한다. 또한 각 문서는 하나 이상의 화제로 구성되며, 문서마다 각 화제와 관련된 분포 비율도 다르다. 토픽 모델링은 핵심어의 화제별 분표 비율과 텍스트의 화제별 분표 비율, 화제별 핵심어 목록과 핵심어별 분포 비율 정보를 산출하여 다각적으로 화제 구조를 파악할 수 있게 한다(홍성연·최재원, 2017; 홍정하·최재웅, 2017).

토픽 모델링의 수행 과정은 다음과 같다. 전처리 과정이 끝난 데이터는 토픽 모델링 수행을 위해 먼저 토픽 수를 지정한다. 일반적으로 토픽 수는 범주화가 가장 잘 되었다고 생각하는 토픽 수를 연구자가 판단하여 정하거나 확률 분포의 복잡도(perplexity)[3] 값을 이용한 토픽 수 결정 알고리즘을 사용한다(이원상·손소영, 2015). 그 후 구현 가능한 알고리즘을 적용하여 토픽 수에 따라 내재된 토픽을 추출한다.

현재 대표적으로 사용되는 알고리즘은 LDA(Latent Dirichlet Allocation)이다. 잠재 디리클레 할당[4] 기법인 LDA는 토픽을 추정하는 화제 분석

---

3) 서로 다른 조건을 같은 트레이닝셋에 놓고, 생성된 값에 대해 어떠한 조건의 결과가 더 나은지 평가하는 방법이다. 복잡도는 결과의 만족도가 높을수록 작은 값을 나타낸다. 즉, 다양한 토픽 수를 선택하여 반복한 후, 복잡도 값이 작은 토픽 수를 선택하는 방식으로 LDA에서 최적의 토픽 수를 산출한다.

4) 기존 확률 잠재 의미 분석은 잠재 의미 변수의 조건부 독립성을 가정한 혼합 모형으로 문서(d)의 확률 분포에 대한 확률 모형을 제시하지 못했다. 따라서 더미 변수 d는 모형의 학습 과정에서 훈련 자료 내의 문서에 대한 경험적 확률변수로 새로운 문서 d'에 대해서는 정의 불가하다는 문제점이 있었다. 이러한 문제점을 극복하고자 Blei et al.(2003)은 토픽을 나타내는 잠재변수와 관련된 은닉 확률변수가 디리클레 분포(dirichlet distribution)를

방법으로, 문서를 구성하는 단어들의 분포를 통해 문서에 잠재되어 있는 토픽들의 비율을 각각 추정하는 방법이다(Blei, 2012). 이때 가장 큰 비율을 지닌 토픽이 해당 문서의 대표 토픽이 된다. 토픽분포는 LDA를 따른다는 전제하에, 미리 상정한 토픽 수와 전체 문서에 출현한 단어들의 분포를 통해 토픽별 단어의 출현 가능성은 간접적으로 추정된다. 즉, LDA는 문맥에서 다른 의미를 가진다면 동일한 단어도 분리하거나 통합할 수 있다는 특징이 있다. 따라서 문서의 토픽 분류에 의미론적인 접근이 가능하다(Born et al., 2014).

III. 연구 방법 및 절차

1. 분석 대상 및 도구

응결성 지수와 응집성 점수와의 상관관계를 분석하기 위해 본 연구에서는 고등학교 3학년 학생이 작성한 논설문 중에서 40편을 임의로 수집하여 분석 대상으로 삼았다. 분석 대상으로 수집한 학생 글 40편은 '환경 미화 우수 학급 시상에 관한 자신의 생각을 서술하라.'는 과제에 따라 고등학생들이 작성한 것이다. 고등학교 3학년 학생을 대상으로 삼은 이유는 고등학교 1학년 이후 쓰기 발달이, 고등학교 2학년 이후 읽기 발달이 안정적 단계에 접어든다는 선행연구에 근거하였기 때문이다(가은아, 2011; 김봉순, 2011). 모든 발달이 완성된 이후인 고등학교 3학년을 대상으로 정하여 기초 언어 능력 부족에서 발생할 수 있는 여타의 변인들과 오차를 통제하고자 했다. 고등학생이 작성한 논설문 40편은 학생들의 쓰기 수준이 골고루 반영될 수 있도록 수집이 이루어졌다.

따른다고 가정하는 잠재 디리클레 할당(latent dirichlet allocation)기법을 제안하였다.

고등학생이 작성한 논설문을 분석하기 위해 이 연구에서는 엑소브레인 한국어 언어 분석 툴킷 Ver 2.0을 사용하였다. 소스 코드 편집에는 .json의 접근을 지원하는 소스 코드 에디터 Visual Studio Code를 활용하였다. Visual Studio Code는 여러 가지 컴퓨터 프로그래밍 언어에서 사용할 수 있는 기능을 갖추고 있으며 디버깅 기능도 지원한다. 이 연구에서는 필요한 어휘의 추출을 위해 .json 포맷 파싱이 가능한 Java 언어가 지원되는 org.json 라이브러리를 사용하였다. 토픽 모델링을 실행하기 위해 기계학습을 가능하게 하는 응용 프로그램인 MALLET 토픽 모델링 패키지를 사용하였다. 또한 채점자들의 평가 신뢰도를 검증하기 위해 EduG와 FACETS 프로그램을 사용하여 국어교사 6명의 논설문 채점 결과를 분석하였다.

2. 분석 절차 및 방법

학생들이 손글씨로 작성한 쓰기 과제는 학생 글 채점 및 분석을 위해 워드프로세서로 입력하였다. 워드프로세서로 입력할 때에는 띄어쓰기는 바르게 수정하였으나, 쓰기 양상을 정확하게 측정하기 위해 표기는 오류가 있더라도 수정하지 않았다. 텍스트 분석을 할 때에는 워드프로세서로 입력한 학생 글을 텍스트 파일(.txt)로 변환하여 사용하였다.

변환한 학생 글은 전처리 과정을 거쳤다. 필요한 품사를 태깅하기 위해 코드 편집기인 Visual Studio를 사용하여 학생 글을 N_Doc 구조체로 나타낸 후, Java와 연동 가능한 프로그램 언어 Scala를 사용하여 명사와 형용사를 추출하였다. 명사와 형용사는 문장에서 의미를 구분하는 주된 요소이기 때문이다. 일반적으로 많이 사용하는 R의 KoNLP 패키지는 어미 변화가 심한 동사와 형용사의 기본형을 추출하는 데 어려움이 많다. 이러한 이유로 본 연구에서는 형용사 추출에 보다 적합한 엑소브레인 한국어 언어 분석 툴킷 Ver 2.0을 사용하였다.

응결성은 범주에 따라 나누어 볼 수 있기 때문에 본고에서는 국지적 응

결, 포괄적 응결, 글 전체 응결 장치의 지수를 토픽 모델링으로 산출하기 위해 토픽 모델링 패키지인 MALLET을 실행하였다. MALLET도 다른 토픽 모델링과 마찬가지로 사용 전에 토픽 수를 미리 설정해야 한다. 이 연구에서는 확률분포의 복잡도 값이 자동 산출한대로 11개의 토픽을 기본 토픽 수로 설정하였다. 그러나 오차를 줄이고 내용상 보다 적절한 토픽 도출을 위해 기본 토픽 수만을 사용하지 않고, ±1의 토픽을 추가로 분석하여 가중치를 두어 계산하였다.[5] 관련된 파라미터는 K값 10, 11, 12, Gibbs 샘플링[6] 기법, 10,000회 반복을 설정하였다. 〈표 1〉은 10개의 토픽으로 군집화한 예이다.

〈표 1〉 토픽 수 10 예시

```
〔Number of Topics〕10
〔Topics in Students〕
-       31      0.5397250607636621
-       33      0.5484865958245239
-       25      0.5656450677605721
-       07      0.5726414968792835
-       08      0.5912213767079815              …
〔Words in Topics〕
-topic  0 words 청찬, 돈, 그룹, 결과, 벌금, 글, 지각, 생산, 달러, 나, 물질 …
-topic  1 words 생각, 좋다, 시상, 아이, 노력, 그렇다, 의미, 활동, 경쟁, 모두, 필요 …
-topic  2 words 사람, 일, 말, 상품, 보상, 자신, 이렇다, 만약, 다르다, 이득, 마음, 의욕, 대가 …
-topic  3 words 청소, 미화, 깨끗하다, 교실, 친구, 시간, 없다, 반, 학교, 선생님, 학년, 마음 …
-topic  4 words 인센티브, 도덕, 효과, 경제, 긍정, 부정, 크다, 적다, 지속, 수단, 필요, 영향 …
-topic  5 words 목표, 관심, 결국, 규칙, 상황, 장점, 결론, 단점, 나중, 능률, 찬성, 단합, 성취감 …
-topic  6 words 학생, 참여, 우수, 이유, 방법, 문제, 개인, 중요, 대회, 심사, 적극 …
-topic  7 words 반, 상, 상장, 상금, 다음, 나쁘다, 뿌듯하다, 공평, 허무, 선물 …
-topic  8 words 환경, 학급, 미화, 시상, 우수, 협동, 목적, 수여, 입장, 의지, 쾌적하다 …
-topic  9 words 보상, 제도, 행동, 경우, 이러하다, 사회, 행위, 사용, 부여, 이익, 목적, 바람직하다 …
```

5) Chang, et al.(2009)는 복잡도로 토픽 수를 결정하는 것에 대해, 토픽 수의 분류가 인간이 하는 인지 판단과 일치하지 않을 수 있음을 지적한 바 있다.

6) 깁스 샘플링(Gibbs Sampling: GS)은 마르코프체인 몬테카를로(Markov chain Monte Carlo: McMC)의 한 종류로 주제-단어분포와 주제 분포 산출을 위한 하나의 기법으로 여러 분포로부터 특정 값을 무작위 추출하여 추정값을 산출하는 방법을 말한다. 일반적으로 잠재변수를 추정할 때엔 본고에서 사용한 깁스 샘플링이나 Variational Inference를 주로 사용한다.

토픽을 추출하고 난 후, 국지적 응결성 지수를 구하기 위해 문장별로 토픽 간의 유사도를 계산하여 문장별 응결성 지수를 학생별로 〈표 2〉와 같이 수치화하였다.

〈표 2〉 1번 학생의 문장별 응결성 지수

〔Number of Topics〕 10
〔Topic Changes in Students〕

-	01	01-02	0.257854821235103
-	01	02-03	0.301749893299189
-	01	03-04	0.130303030303030
-	01	04-05	0.234920634920635
-	01	05-06	0.323809523809524
-	01	06-07	0.221457489878543
-	01	07-08	0.287449392712551
-	01	08-09	0.184615384615385
-	01	09-10	0.243724696356275

앞의 01은 학생 번호를 의미하며, 뒤의 01-02는 문장 번호이다. 즉, 첫 줄의 0.257854821235103은 1번 학생의 첫 번째 문장과 두 번째 문장 사이의 응결성 지수를 나타낸다. 〈표 2〉와 동일한 방법으로 토픽의 수를 달리 하여 반복 군집화(clustering) 한 후, 결과 값에 가중치를 주어 학생 글의 국지적 응결성 지수를 산출하였다.

포괄적 응결성 지수는 문단 간의 유사도로 측정하였는데, 방법은 국지적 응결성 지수 산출과 동일하다. 그러나 국지적 응결성과 달리, 한 개의 문단을 하나의 문서로 설정하였다는 점이 다르다.

글 전체의 응결성은 설계를 달리하여 산출하였는데, 기계 학습의 한 범주인 준 지도 학습(Semi-Supervised Learning)[7] 방식을 활용하였다. 먼저

---

7) 준 지도 학습은 목표 값이 충분히 표시된 학습 데이터를 사용하는 지도 학습(Supervised Learning)과 목표 값이 표시 되지 않은 데이터를 사용하는 자율 학습(Unsupervised Learning) 사이에 위치한다. 일반적으로 이 방법은 목표 값이 표시된 학습 데이터가 적고 표시되지 않은 데이터를 많이 갖고 있을 때 사용한다. 준 지도 학습은 지도 학습으로 목표 값을 산출한 후 비 지도 학습으로 자동으로 목표 값을 산출하게 하여 반복 학습을 하는 방

작문 교과 전문가 2인에게 의뢰해 상, 중, 하의 기준이 되는 학생 글을 각각 세 편씩 선정하였다. 학습 데이터[8]를 이용한 학습을 위해서 자질(feature)을 추출한 후[9] 자질 가중치 학습을 반복하여 결과 값을 산출했다. 〈표 3〉은 전문가가 부여한 점수를 학습시켜 최적의 가중치 값을 찾은 결과이다.

〈표 3〉 학습 데이터의 목표 값 산출

```
〔Topic Weights〕
- 0.13342392188771293 -0.2270448425525808
  0.01594128260159161 …
〔Reward Sum〕 7.6531678509037615
〔Best Reward Sum〕 8.501866584686569
〔Topic Changes in Students〕 …
〔Sum〔Topic Changes〕 in Students〕
- 07 4.12990469336437443
- 08 2.83924760598411809
- 18 4.08799985419210186
- 25 0.91786770424366452
- 27 1.26273804020796108
- 29 3.98231084093336354
- 30 0.99394187264114994
- 37 2.92199428026425936
- 39 3.27296327915210298
```

자질 가중치 학습 후 미분류 집단을 군집화하여 결과 값을 도출하였는데, 이때 확률 분포는 가우시안 분포를 따랐다.

한편, 학생 글은 현재 석사 과정에 있는 현직 교사 6명이 채점하였다. 평가 기준표는 연구의 목적에 맞게 박영민·김승희(2007)의 기준표를 응집성 항목이 포함되도록 변형하여 활용하였다. 평가 기간은 2017년 7월 25일부터 8월 5일까지였다. 본 연구에서 글 전체의 평가 결과는 논의의 대상이 아

법이다. 이 방법은 지도 학습에 비해 비용이 적게 들고, 자율 학습에 비해 정확도가 높다는 장점이 있어 최근 여러 분야에서 사용되고 있다(Sogaard, 2013).

8) 〈표 3〉에서 보는 바와 같이, 학습 데이터는 7, 8, 18, 25, 27, 29, 30, 37, 39번 학생 글로 하였다.

9) 기계 학습을 위해서는 문제를 해결하는 데 도움이 되는 자질을 선정하여 학습을 하는 것이 중요하다. 본 연구에서는 명사와 형용사를 내용어로 상정하고, 단어 자질을 추출하였다.

니었으므로, 평가 결과 중 응집성 항목 점수만을 사용하였다. 응집성에 대한 평가는 평가자들의 총체적 평가로 진행하였다. 평가자 간 신뢰도를 확인하기 위해 EduG 6.1을 사용하여 일반화 가능도 값을 산출하였다. 일반화 가능도 계수는 0.79로 평가자들은 신뢰할 만한 수준을 보였다. 그런데, 학생(S)에 대한 평가자(R)들의 평가가 상이하여 오차의 크기가 컸기 때문에, 평가자별로 부적격 평가자가 있는지 추가로 확인하는 작업이 필요했다. 이에 FACETS을 사용하여 부적합, 과적합 평가자를 확인하는 절차를 진행했다.

IV. 연구 결과 및 논의

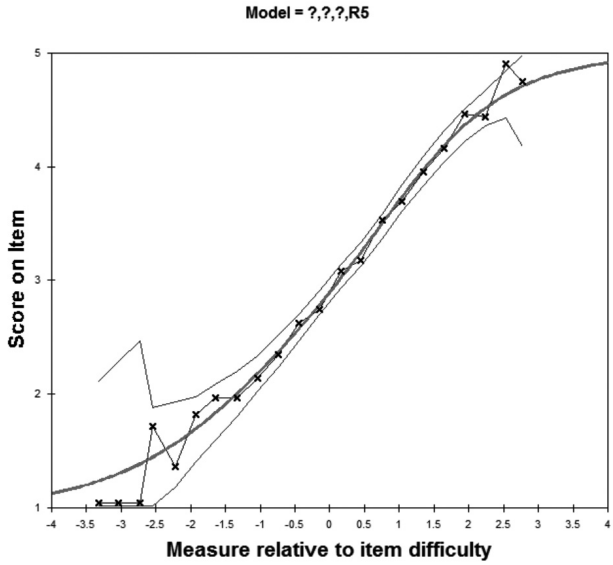
1. 평가자 응집성 채점 결과

교사 6명의 응집성 항목에 대한 채점 결과의 기술통계는 〈표 4〉와 같다.

〈표 4〉 응집성 채점 결과 기술통계

	표본크기	최소값	최대값	평균	표준 편차
응집성 점수	40	1.83	4.50	3.3788	.62001

다음으로, FACETS을 사용하여 부적합, 과적합 평가자를 확인하기 위해 우선, 모형 적합도 검증 그래프를 통해 평가 결과가 다국면 Rasch 모형에 적합한지 분석하였다. 〈그림 1〉에 따르면 교사 평가자들의 응집성 평가 결과는 분석 모형에 적합한 것으로 나타났다.



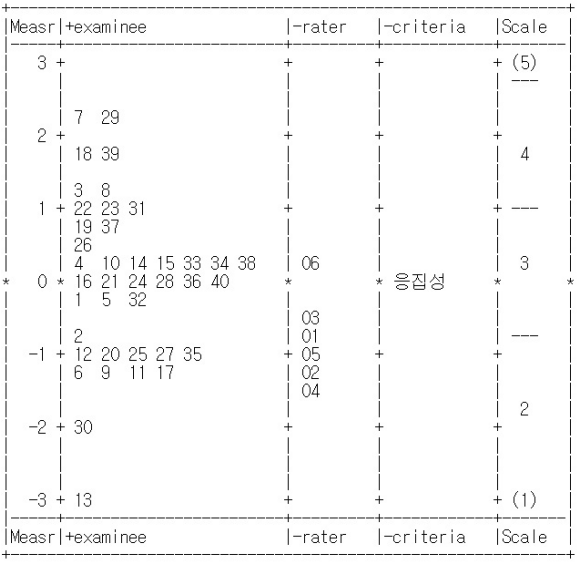
〈그림 1〉 모형 적합도 분석 결과

교사 평가자들의 응집성 평가 자료를 FACETS 프로그램으로 분석하여 얻은 logit 값을 도식화하면 〈표 5〉와 같다.

〈표 5〉 평가자 적합도

평가자	관찰값	기댓값	logit 값	표준오차	내적합도	내적합지수	외적합도	외적합지수
4	3.80	3.86	−1.48	.22	.62	−1.9	.62	−1.9
2	3.63	3.67	−1.16	.21	1.49	2.0	1.40	1.7
5	3.47	3.51	−.90	.21	.86	−.6	.86	−.6
1	3.38	3.40	−.73	.21	1.06	.3	1.02	.1
3	3.26	3.26	−.52	.21	1.00	.0	.98	.0
6	2.75	2.70	.34	.21	.99	.0	.96	−.1

평가자의 적합도는 내적합제곱평균과 내적합표준화 값으로 판단 가능하다. 분석 결과 2번 채점자의 내적합 지수는 다소 과적합 양상을, 4번 채점자의 내적합 지수는 다소 부적합 양상을 보였다. 그러나 내적합 지수가



〈그림 2〉 FACETS 분석 요약

−2.0와 2.0 사이에 있었으므로 2번, 4번 채점자의 평가 결과를 분석에서 배제하지 않았다.[10] 학생 글에 대한 여섯 평가자들의 평가 결과는 〈그림 2〉와 같았다.

7번과 29번 학생이 가장 높은 점수를, 13번 학생이 가장 낮은 점수를 받았고, 학생들의 점수는 대체로 정규분포와 같은 분포도를 보였다.

2. 응결성 지수와 응집성의 상관관계

1) 국지적 응결성 지수와 응집성의 상관관계

문장 간 유사한 정도가 높으면 국지적 연결성 지수가 높게 드러난다. 이에, 문장 간 유사한 정도에 대해 토픽 모델링으로 산출한 점수를 국지적 응

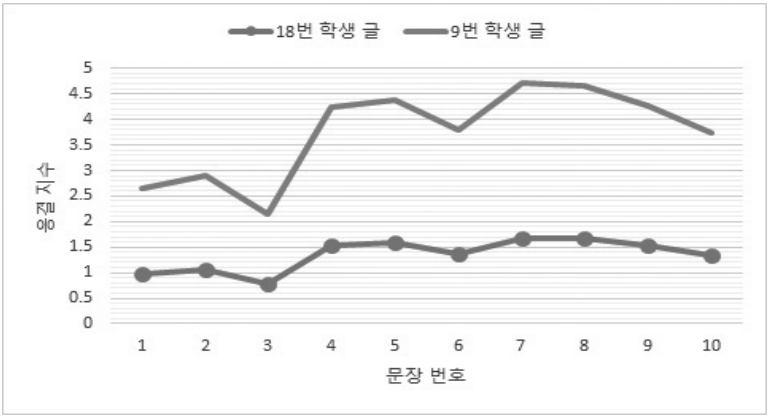
10) 일반적으로 0.75에서 1.3 logit 사이에 분포하는 측정값을 적정값으로 판단하나, 허용적으로 0.5에서 1.5 logit 사이에 위치하는 값을 적합한 값으로 간주하기도 한다.

결성 지수[11]로 설정하여 응집성과의 상관관계를 구하였다.

〈표 6〉 응집성 점수와 국지적 응결성 지수의 상관관계

		국지적 응결성
응집성	Pearson 상관계수	-.634
	유의수준(양쪽)	.000

〈표 6〉에서 보는 바와 같이 국지적 응결성 지수는 채점 결과와 부적 상관관계를 나타냈다. 특히 부적 상관계수가 높음에 주목할 필요가 있는데, 이는 문장 간 토픽의 중복을 능숙한 독자인 교사 평가자들은 오히려 응집성이 부족한 글로 인식하고 있다는 것이기 때문이다. 이러한 결과는 학생들이 사용한 중복되는 단어와 개념이라는 응결 장치가 글의 내적 의미망을 형성하는 데에 영향을 미치지 못했음을 시사한다. 〈그림 3〉은 응결성 지수의 경향성을 살펴보기 위해 응집성 점수가 상위이면서 국지적 응결성 지수가 하위인 18번 학생 글과 응집성 점수가 하위이면서 국지적 응결성 지수가 상위인



〈그림 3〉 9번, 18번 학생 글의 문장별 응결성 지수

11) 토픽 모델링으로 산출한 결과 값은 최댓값 4.37, 최솟값 0.78이었다.

9번 학생 글의 문장별 응결성 지수를 나타낸 것이다.

두 학생 모두 서론에 비해 본론 부분에서 상대적으로 더 높은 응결성을 보였다. 또한 9번 학생 글은 18번 학생의 글에 비해 모든 문장의 응결성 지수가 더 높았다. 문장의 응결성 지수가 높은 학생일수록 글을 쓸 때, 새로운 의미를 생성해 내는 데 어려움을 겪고, 대부분의 문장에서 앞 문장과 유사한 내용을 반복 제시하였다.

〈학생 글 9번 4-7 문장〉

경제적 인센티브는 또 부정적, 긍정적으로 나누어진다. '환경 미화 우수 학급 시상'에 관해서는 긍정적 인센티브에 해당된다. '환경 미화 우수 학급 시상'은 좋은 일에 대한 보상을 줌으로써 환경 미화에 관심 없는 사람도 관심을 가지게 된다. 좋은 행위를 더 많이 하기 위해서는 보상이나 이익이 증가해야한다.

〈학생 글 18번 4-7번 문장〉

이를 논하기 전에, 시상에 대해 먼저 언급할 필요가 있다. 시상이라는 인센티브가 필요한가, 불필요한가? 얼마 전 본 신문에는 '인센티브가 우리의 도덕 감정을 몰아낼 수 있지만 반대로 이러한 인센티브가 성공을 거둔 경우가 있다'고 적혀 있었다. 어떠한 도덕적 감정에 더 큰 명예와 더 많은 권한을 부여하는 것이 중요하다는 것이다.

위의 두 예시문은 국지적 응결성 지수와 응집성이 부적 상관관계를 보이는 이유를 짐작케 하는 자료이다. 9번 학생의 글은 문장 간 의미의 반복이 두드러지며, 함축하고 있는 내용이 매우 유사하다. 그러나 18번 학생의 글은 다음 문장에서 앞 문장의 내용을 일부 받으면서도 논의를 확장하고 있음을 알 수 있다.

2) 포괄적 응결성 지수와 응집성의 상관관계

전술하였듯이, 포괄적 응결성은 텍스트의 단락 간 의미적 유사성을 통해 확인할 수 있다. 문단 간의 유사한 정도를 포괄적 응결성으로 설정하고, 응집성 점수와 상관관계를 산출하였다.

〈표 7〉 응집성 점수와 포괄적 응결성 지수의 상관관계

		포괄적 응결성
응집성	Pearson 상관계수	.289
	유의수준(양쪽)	.070

〈표 7〉의 결과처럼 포괄적 응결성 지수는 응집성 점수와 0.289로 양의 상관관계를 나타내었다. 국지적 응결성 지수에서 음수의 결과가 나왔던 양상과는 사뭇 다른데, 이는 교사 평가자들이 문단 간에 유사한 토픽이 반복되는 글은 문장 간 반복에 비해 부정적으로 인식하지 않는다는 것을 의미한다. 그러나 이것이 긍정적으로 인식했다는 것을 함의하는 것은 아니다. 통계적으로 그 값이 유의하지 않았기 때문이다. 이는 포괄적 응결성 지수와 응집성 사이에 선형적인 관련성이 존재하지 않음을 의미한다. 즉, 포괄적 응결성 지수가 높은 글이 반드시 응집성이 좋은 글이라고 말할 수는 없다는 것이다. 이는 평가자들이 생각하는 응집성이 좋은 글은, 문단 간 유사한 토픽이 유기적 연결 관계를 맺고 있는 글일 수도 있지만, 그보다 문단이 전개될수록 충분한 내용의 심화와 논의의 확장이 일어나는 글이라는 점을 짐작할 수 있다. 일반적으로 응결성 지수를 측정하는 여러 연구에서 포괄적 응결성을 사용하는데, 그 타당성에 대한 재고가 필요하다. 또한, 이는 최근 국외를 중심으로 연구되고 있는, 자동 응결성 분석 도구인 TAACO(the Tool of the Automatic Analysis of Text Cohesion)의 유용성에 대해서도 비판적 검토가 필요함을 시사하는 대목이라 할만하다.[12)]

12) Crossley et al.(2015)은 TAACO를 이용하여 글을 측정한 결과 포괄적 응결성 지수가 글

3) 글 전체 응결성 지수와 응집성의 상관관계

글 전체 응결성은 학습 데이터를 바탕으로 글 전반에 사용된 토픽의 유사도를 수치로 산출하였다. 〈표 8〉은 글 전체의 응결성 지수와 응집성과의 상관관계를 살펴 본 결과이다.

〈표 8〉 응집성 점수와 글 전체 응결성 지수의 상관관계

		글 전체 응결성
응집성	Pearson 상관계수	.854
	유의수준(양쪽)	.000

글 전체 응결성 지수는 응집성 점수와 높은 상관관계를 나타냈고, 통계적으로도 그 결과가 유의했다. 두 점수의 절대적 지수를 단순히 비교하기보다 교사 채점자들이 생각하는 응집성이 높은 글과 글 전체 응결성 지수가 높은 글이 얼마나 유사한지를 살피기 위해 평가자 채점 순위와 글 전체 응결성 순위를 〈표 9〉에서 비교해 보았다.

〈표 9〉 학생 글 평가 순위

학생 글 번호	평가자 응집성 채점 순위	글 전체 응결성 순위	학생 글 번호	평가자 응집성 채점 순위	글 전체 응결성 순위
1	26	31	21	20	34
2	29	33	22	7	9
3	5	6	23	7	29
4	13	21	24	20	14
5	26	19	25	30	39
6	35	23	26	12	10
7	1	1	27	30	30
8	5	8	28	20	12

의 응집성과 양의 상관관계를 보이고, 글 전체의 질과도 양의 상관관계를 보인다고 밝힌 바 있다.

9	35	32	29	1	3
10	13	16	30	39	36
11	35	24	31	7	4
12	30	22	32	26	14
13	40	40	33	13	11
14	13	25	34	13	27
15	13	28	35	30	26
16	20	35	36	20	20
17	35	38	37	10	7
18	3	2	38	13	12
19	10	17	39	3	5
20	30	37	40	20	17

두 순위 간에 비교적 큰 차이를 보이는 글도 있었지만, 대체로 교사 평가자들이 생각하는 순위와 글 전체 응결성 순위가 비슷한 글들이 많았다. 이는 학생 글 수준에 따라 사용하는 의미역이 유사한 경향을 보인다는 것을 방증하는 예라고 볼 수 있다. 아래의 글은 이러한 경향성을 확인해 볼 수 있는 예시이다.

〈학생 글 30번 일부〉

근데 이건 담임선생님의 역할이 좀 크다. 평소에도 좀 깨끗하신 선생님이면 환경 미화에 엄청 신경 쓰신다. 그런데 그런 것을 별로 신경 안 쓰시는 선생님이면 환경 미화 검사 날이 언제인지 기억도 안 날 정도로 그냥 넘어가 버린다.

이걸 보면 사람의 관점마다 다르다는 생각이 들었다. 학교에서 이 대회를 진행한 이유는 1년에 한 번은 청소를 열심히 하자는 것일 텐데 할 사람은 하고 안 할 사람은 안하니 인센티브가 안 먹힌 것 같다.

〈학생 글 13번 일부〉

보상을 받은 학급들은 당일에만 효과를 가지고 추후 효율성이 떨어질 것이다. 그러나 상품이 칭찬이란 걸 알면 경쟁이 일어나지 않겠지만 환경 미화에도 신

경을 쓰지 않을 것이다. 그렇다고 우수하지 않은 학급이 벌금을 내면 불만이 많이 생성되면서 억지로 청소할 것이다.

나는 환경 미화 우수 학급 시상에 별로 관심이 없다. 청소도 즐겁게 하는 것이 좋고 상품 받는 기분이 배가 될 텐데 하고 싶지 않은 친구들이 있으면 하면서도 불만을 표현하니까 효율성이 떨어진다.

위에 제시된 30번 학생 글은 전문가 2인이 학습 데이터 채점 시 낮은 점수로 책정한 글이고, 13번 학생 글은 미분류 집단 글로, 프로그램에 의해 점수가 자동 산출된 글이다. 이 두 글은 6명의 교사 채점자들에게 최하위 점수를 받았는데, 글 전체 응결성 지수를 산출한 결과 30번 글은 0.99을 13번 글은 0.91로 유사한 정도의 낮은 점수를 받았다. 교사 채점자가 총체적 관점에서 응집성이 낮은 글이라고 평가한 글들에는 유사한 경향이 존재했다.

두 글에는 모두 응집성을 해치는 문장이 포함되어 있고, 중심 의미의 결속력도 낮았다. 이러한 특징은 평가자들의 인식뿐 아니라, 프로그램으로 산출한 글 전체의 응결성 지수에도 수치화되어 나타났다. 환원하면, 학생 글의 수준에 따라 글 전체에 드러나는 토픽의 유사도에 일정한 경향성이 존재하고, 이러한 경향성은 프로그램의 학습 알고리즘을 통해 자동으로 수치화가 가능하다는 것이다. 이 지점에 착목할 필요가 있는데, 이는 글 전체에서 토픽들 간의 유사도 비율을 산출해 내는 방법이 글의 응집성을 평가하는 새로운 방안이 될 수 있음을 시사하기 때문이다.

V. 맺음말

이 연구에서는 토픽 모델링을 사용하여 산출한 학생 글의 응결성 지수와 응집성 점수 간에 어떤 상관관계가 있는지에 대해 분석하였다. 이를 위해

응결성을 국지적, 포괄적, 글 전체 응결성으로 구분한 후 토픽 유사도를 바탕으로 각각의 응결 지수를 분석하였고, 그 결과 값을 교사 채점자들이 총체적으로 평가한 응집성 점수와 비교하였다. 그 결과, 응결성의 범주에 따라 교사 채점자의 응집성 점수와 상이한 상관관계를 보였다.

지금까지 이 글에서 논의한 바를 약론하면 첫째, 국지적 응결성 지수와 교사 채점 결과는 부적 상관관계를 나타냈다. 문장 간 토픽 유사도가 높다는 것은 앞·뒤 문장 간의 단순한 의미 반복이 두드러지기 때문이며, 교사 평가자들은 이를 내용 생성의 어려움에서 기인한 것으로 판단하여 응집성이 낮은 글이라 평가한 것이다. 둘째, 포괄적 응결성 지수와 교사 채점 결과는 0.289였으나 선형적인 관련성이 존재하지 않았다. 응결성과 응집성의 상관관계를 측정하는 여타 연구에서 응결성 지수의 대푯값으로 포괄적 응결성 지수를 측정하고 있는데, 그 타당성에 대한 검증이 선행되어야 할 것이다. 셋째, 글 전체 응결성 지수와 교사 채점 결과는 정적 상관관계를 보였고, 통계적으로도 유의했다. 교사 채점자가 인식하는 응집성 수준과 글 전체 토픽의 유사도에는 특정한 경향성이 존재했고, 이러한 이유로 토픽 모델링을 활용하여 교사 채점자와 유의한 결과의 도출이 가능했다.

즉, 응결성과 응집성의 관계는 응결성을 어떤 범주에서 측정하느냐에 따라 그 결과에 차이가 있었다. 이는 박혜진·이미혜(2017)에서 응결 장치에 따라 응집성과의 관계가 상이했다는 결과와도 일정 부분 맥을 같이 하는바, 앞으로 응결성에 관한 연구에서 범주에 따른 분석도 필요함을 밝히는 대목이라 하겠다.

토픽 모델링을 활용하여 분석한 응결성 지수 중 글 전체 응결성 지수가 교사 채점자들의 응집성 지수와 유의하다는 앞선 논의는, 텍스트 응결성이 글 전체의 토픽 유사도라는 응결 장치를 통해 드러나고 이를 통해 텍스트의 응집성을 파악할 수 있음을 말해 준다. 또한 토픽 모델링이라는 자동 산출 방법이 글에 대한 객관적 정보를 제공하는 타당한 도구가 될 수 있음을 방증한다.

국어교육에서의 텍스트 분석에 대한 본격적 연구는 이제 시발점에 들어섰다. 기술의 발달로 추동된 텍스트 분석은 작문 분야에서뿐만 아니라 독서, 문학 영역에까지 경계를 허물고 확장될 수 있는 또 하나의 방법론으로 자리 잡을 수 있을 것이다. 그러나 이러한 연구들이 발전적 방향으로 나아가기 위해서는 방법상의 고도화·정교화 방안과 더불어 타당성에 대한 비판도 지속적으로 제기되어야 할 것이다. 특히 양적으로 수치화되는 지표들이 보여줄 수 있는 교육적 함의에 대한 의문은 연구의 올바른 방향성을 제시해 주는 동시에 반성적 기회를 제공해 줄 수 있을 것이다.

* 본 논문은 2017. 8. 12. 투고되었으며, 2017. 8. 21. 심사가 시작되어 2017. 9. 11. 심사가 종료되었음.

참고문헌

가은아(2011),「쓰기 발달의 양상과 특성 연구」, 한국교원대학교 박사학위논문.

김민영(2014),「설득 텍스트의 결속구조 표지 양상 연구: 초등학교 6학년을 대상으로」, 한국교원대학교 석사학위논문.

김봉순(2011),「아동기와 청소년기의 문식성 발달」,『공주교대논총』46(2), 25-56, 공주교육대학교 교육학술정보원.

김정숙(1996),「담화능력 배양을 위한 읽기 교육 방안」,『한국말 교육』7, 295-309, 국제한국어교육학회.

박영민(2004),「국어과 교육과정 용서의 진술과 개념: 통일성, 응집성, 일관성을 중심으로」,『독서연구』11, 181-206, 한국독서학회.

박영민·김승희(2007),「쓰기 효능감 및 성별 차이가 중학생의 쓰기 수행에 미치는 효과」,『국어교육학연구』28, 327-360, 국어교육학회.

박영순(2008),『한국어 담화·텍스트론』, 한국문화사.

박진용(2003),「읽기 교수·학습을 위한 텍스트구조의 전개과정 고찰」,『청람어문교육』27, 1-25, 청람어문교육학회.

박혜진·이미혜(2017),「한국어 학습자의 쓰기 텍스트에 나타난 응결성과 응집성의 상관분석」,『우리말글』73, 133-157, 우리말글학회.

서승아(2008),「7학년 쓰기 능력의 응집성 발달특성 추출: 설득적 텍스트의 응집성에 대한 학생 상호 첨삭방법의 활용」,『국어교육』126, 157-184, 한국어교육학회.

이신형(2012),「화제 중심의 텍스트 응집성 고찰」,『청람어문교육』45, 297-323, 청람어문교육학회.

이원상·손소영(2015),「공간빅테이터 연구 동향 파악을 위한 토픽모형 분석」,『대한산업공학회지』41(1), 64-73, 대한산업공학회.

정다운(2008),「한국어 학습자의 작문에서 '내용 조직하기' 분석: 응집성을 중심으로」,『어문연구』139, 419-442, 한국어문교육연구회.

한국텍스트언어학회(2004),『텍스트 언어학의 이해』, 박이정.

홍성연·최재원(2017),「토픽 모델링 분석 기법을 활용한 대학의 학생 지원 연구 동향 분석」,『학습자중심교과교육연구』17(8), 21-48, 학습자중심교과교육학회.

홍정하·최재웅(2017),「토픽 모델링을 이용한 코퍼스의 주제구조 탐색」,『언어와 정보 사회』30, 239-275, 서강대학교언어정보연구소.

Beaugrande, R. De., & Dressler, W. U.(2008),『텍스트 언어학 입문』, 김태옥·이현호 역, 한신문화사(원서출판 1981).

Blei, D. M.(2012), "Probabilistic topic models," *Communications of the ACM* 55(4), 77-84.

Born, J., Scheihing, E., Guerra, J., & Carcamo, L.(2014), "Analysing microblogs of mid-

dle and high school students," *European Conference on Technology Enhanced Learning*, 15-28.

Brinker, K.(1988), *Linguistische Textanalyse*, Berlin: Erich Schmidt Verlag.

Chang, J., Gerrish, S., & Wang, C.(2009), "Reading tea leaves: How humans interpret topic models," *Advances in Neural*, 288-296.

Crossley, S. A., & McNamara, D. S.(2012), "Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty," *Journal of Research in Reading* 35, 115-135.

Crossley, S. A., Kyle. K.,& McNamara, D. S.(2015), "The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion," *Behavior Research Methods* 35, 115-135.

Halliday, M. A. K., & Hasan, R.(1976), *Cohesion in English*, London: Longman.

Sogaard, A.(2013), "Semi-Supervised Learning and Domain Adaptation in Natural Language Processing," *Morgan & Claypool Publishers* 6(2), 519-522.

부록

학생 글 번호	평가자 응집성 점수	글 전체 응결성 지수	학생 글 번호	평가자 응집성 점수	글 전체 응결성 지수
1	3.25	1.22	21	3.44	1.08
2	3.16	1.09	22	3.96	2.74
3	4.14	3.01	23	3.96	1.31
4	3.63	1.77	24	3.44	2.20
5	3.25	1.90	25	3.13	0.92
6	3.02	1.70	26	3.73	2.63
7	4.57	4.13	27	3.13	1.26
8	4.14	2.84	28	3.44	2.26
9	3.02	1.11	29	4.85	3.98
10	3.63	2.04	30	2.81	0.99
11	3.02	1.58	31	3.96	3.71
12	3.13	1.72	32	3.25	2.20
13	2.77	0.91	33	3.63	2.36
14	3.63	1.52	34	3.63	1.47
15	3.63	1.32	35	3.13	1.50
16	3.44	1.04	36	3.44	1.87
17	3.02	0.94	37	3.82	2.92
18	4.49	4.09	38	3.63	2.26
19	3.82	1.98	39	4.49	3.27
20	3.13	0.98	40	3.44	1.98

초록

토픽 모델링에 따른 고등학생 논설문의 응결성과 응집성의 상관분석

이슬기

이 연구의 목적은 토픽 모델링을 사용하여 산출한 고등학생 논설문의 응결성 지수와 응집성의 상관관계를 분석하고, 응결성 측정 방법에 대한 타당성을 고찰하는 데 있다. 이를 위해 응결성을 국지적, 포괄적, 글 전체 응집성으로 세분한 후 토픽 유사도로 응결 지수를 분석하였고, 결과 값을 교사 채점자들의 응집성 점수와 비교하였다. 연구 결과, 국지적 응결성 지수는 부적 상관관계에 있었는데, 문장 간의 유사도가 높은 글을 교사 평가자들은 내용 생성의 어려움을 겪는 글로 평가했다. 포괄적 응결성 지수는 교사 채점 결과와 유의하지 않았고, 글 전체 응결성 지수는 정적 상관관계를 보였다. 이는 교사 채점자가 인식하는 응집성 수준에 따라 글 전체에 드러나는 토픽의 유사도에 특정한 경향성이 존재한다는 것을 의미한다. 이 연구를 통해 응집성은 글 전체의 응결성과 유관하고, 토픽 모델링이 글에 대한 객관적 정보를 제공하는 타당한 도구가 될 수 있음을 확인할 수 있었다.

핵심어 응결성, 응집성, 응결 장치, 토픽 모델링, LDA, 기계 학습, 준 지도 학습, 텍스트 마이닝, RASCH 모형, 일반화 가능도

ABSTRACT

A Correlation Analysis of Coherence and Cohesion in High School Students' Essays through Topic Modeling

Lee Seulki

This study investigates the validity of the method used to measure coherence by analyzing the correlation between cohesion scores and the coherence of student texts calculated using topic modeling. For this purpose, cohesion through topic similarity was categorized into local, global, and overall text coherence. Next, the results were compared with teacher-generated coherence scores. The results showed that local cohesion was negatively correlated with the coherence scores. Global cohesion was not significantly correlated with teachers' grading results, and overall text cohesion showed a positive correlation. This study shows that coherence is related to the consistency of the whole text and that topic modeling can be a useful tool for providing objective information about the text.

KEYWORDS Cohesion, Coherence, Cohesive Device, Topic Modeling, LDA, Machine Learning, Semi-Supervised Learning, Text Mining, Rasch Model, Generalizability