

# 한국어 텍스트 결속성 측정 도구: KorCAT

김동현<sup>1</sup>, 김현정<sup>2</sup>, 김철희<sup>2</sup>, 서영덕<sup>1,†</sup>

인하대학교 컴퓨터공학과<sup>1</sup>, 인하대학교 국어교육과<sup>2</sup>  
{gmr4rnjs, Kimbean0324}@gmail.com, {clearfe, mysid88}@inha.ac.kr

## KorCAT: Cohesion Analysis Tool for Korean Text

Dong-Hyun Kim<sup>1</sup>, Hyun-Jung Kim<sup>2</sup>, Chul-hui Kim<sup>2</sup>, Young-Duk Seo<sup>1,†</sup>  
Department of Computer Science and Engineering, Inha University<sup>1</sup>,  
Department of Korean Language Education, Inha University<sup>2</sup>

### 요약

결속성은 텍스트의 의미 관계 분석에서 주요하게 고려되는 요인 중 하나이다. 결속성이 높을 수록 텍스트 내 언어적 의미 관계가 긴밀한 것으로 볼 수 있기 때문이다. 사람에 의한 텍스트 결속성 분석은 주관성이 개입되는 문제를 배제할 수 없는데, 영어에는 TAACO라고 불리는 결속성 자동 측정 도구가 있어 다양한 지표들로 텍스트의 결속성을 측정하고 있다. 그러나 한국어에는 이와 같은 텍스트 결속성 자동 측정 도구가 거의 없으며, 관련 연구도 상당히 부족한 실정이다. 따라서 본 논문에서는 한국어 텍스트의 의미 관계 분석을 위한 결속성 지표를 정의하고, 한국어의 특성에 맞는 결속성 측정 도구인 KorCAT을 제안한다.

**주제어:** 텍스트 결속성, 구조적 결속성, 의미적 결속성

### 1. 서론

언어학에서 텍스트의 개념은 매우 다양하게 정의되어 왔으며, 텍스트의 성격을 규정하는 연구도 꾸준히 이루어졌다. 대표적으로 Beaugrande와 Dressler은 텍스트를 텍스트로서 존재하게 하는 7가지 텍스트성(Textuality)을 제시하였다 [1]. 결속성(Cohesion)과 응집성(Coherence)은 7가지 텍스트성 중 가장 중요하게 여겨진다. 결속성은 문장과 문장 혹은 단락과 단락을 연결하여 텍스트를 구조적으로 긴밀하게 연결하는 자질을 의미하며, 응집성은 텍스트 전체의 의미적 연결성을 가리킨다 [2].

응집성과 결속성의 용어사용과 관계를 논의하는 연구는 다수 존재해왔으며 학자들마다 상이한 양상을 보인다. 그 중 [3]은 응집성과 결속성의 관계를 크게 세 가지로 나눈다. 결속성이 응집성을 갖게 하는 중요한 요소라는 입장과 별다른 도움이 되지 않는다는 입장, 서로 상호작용한다는 입장이 그것이다. 이처럼 한국어 교육에서 연구자마다 다른 용어를 사용하고 다른 입장을 취하지만 많은 학자들이 결속구조를 통한 텍스트 표층의 의미연결이 텍스트의 의미연결로 이어진다고 본다. 즉 텍스트 표층의 다양한 문법적 결속장치로 단어, 구, 절, 문장이 연결되고 문단이 이어져 완결된 텍스트로 엮어지는 특성이 결속성이며 결속성이 통일된 텍스트의 내용적 일관성에 기여한다는 점에 동의한다는 것이다 [2].

작문은 필자의 의미를 재구성하는 행위임과 동시에

다양한 결속기제를 활용하여 언어 요소들 간의 의미를 긴밀하게 조직하여 자신의 의도를 전달하는 과정이라고 볼 수 있다 [4]. 때문에 한국어 작문 교육에서는 이를 고려하여 의미적으로 일관성이 있으며 논리적으로 완결성이 있는 글을 생산하는 것을 목표로 하기도 한다. 능숙한 필자의 경우 문장을 연결하기 위해 적절히 결속기제를 사용하여 의미관계가 긴밀한 글을 생산할 수 있는 능력이 있다는 연구 결과들이 다수 존재한다 [3]. 이를 근거로 결속성과 응집성은 전통적으로 작문 평가의 중요한 요소로 여겨져 왔다. 하지만 텍스트의 결속성은 결속기제에 의해서만 구성되는 것이 아니라, 필자와 독자의 배경지식, 의도, 맥락과 문화적 요인 등과 관련되어 있어 객관적 지표로 제시되는 데 한계가 존재했다 [2]. 따라서 지금까지 이러한 한국어 텍스트 결속성의 분석과 평가는 전적으로 인간 평가자의 전문성과 주관성에 의존해 왔다.

2028년에 도입을 검토중인 대학수학능력시험의 논술·서술형 문항 평가[5]나 한국어능력시험(TOPIK) 작문 영역 평가와 같은 대규모 작문 평가를 위해서는 이러한 평가자의 전문성과 주관성 문제 외에도 평가에 많은 시간과 비용이 드는 문제 역시 해결해야 한다. 영어의 경우, 이러한 대규모 작문 평가에 TACCO와 같은 텍스트 의미관계 자동 분석 프로그램을 개발하여 적극적으로 활용하고 있다. 그러나 한국어에서는 한국어 텍스트의 의미관계를 자동적으로

† 교신저자 (Corresponding author)

로 분석해 주는 도구 개발이 거의 이루어지지 않았으며, 관련 연구도 상당히 부족한 실정이다. 현재까지 개발된 한국어 결속성 분석 도구는 형태소 분석 과정 없이 대명사, 접속사, 단어 빈도수만 측정하기 때문에 세밀하고 정확한 결속성 분석에 한계가 있다 [5]. 따라서 본 논문에서는 영어 텍스트의 결속성 분석 프로그램인 TACCO의 개발 사례를 참고하되, 형태소 단위를 적극적으로 고려해야 하는 한국어의 특성을 반영하여 한국어 텍스트 분석에 적합한 결속성 자동 분석 도구 KorCAT (Cohesion Analysis Tool for Korean text)를 개발 및 제안하고자 한다.

이를 위해 먼저, 한국어 텍스트의 결속성을 구조적 결속성(Structural cohesion, Cohesion)과, 의미적 결속성(Semantic cohesion, Coherence)으로 구분하고, 이를 계량할 수 있는 지표를 구안하고자 한다. 구조적 결속성은 접속사, 대명사 등의 문법 요소 빈도 계량을 통해 문장과 문장이 구조적으로 긴밀하게 연결되었는지를 분석할 수 있다. 의미적 결속성은 기존 한국어 결속성 측정 도구의 결점을 보완하기 위한 지표로 문장 간 의미적인 연결성을 측정할 수 있는데, Top2Vec을 사용하여 주제간 응집도를 추출하고 SBERT를 활용하여 문장 간 유사도를 평가할 수 있다.

본 논문의 구조는 다음과 같다. 2장에서는 결속성 측정에 관한 기존 연구에 대해 소개한다. 3장에서는 한국어 텍스트의 결속성을 측정하는 방안을 제시하며, 본 논문에서 제안하는 KorCAT을 자세히 설명한다. 4장에서는 KorCAT을 사용하여 초, 중, 고 교과서 텍스트의 구조적, 의미적 결속성을 측정한 결과를 제시한다. 마지막으로 5장에서는 본 논문을 정리하고 향후 연구방향을 제시한다.

## 2. 관련 연구

### 2.1 국내의 결속성 자동 측정 연구

한국어는 교착어이기 때문에 영어를 비롯한 타언어에 비해 결속성 측정이 어려워 결속성 측정 도구가 많지 않다 [6, 7]. [7]는 영어 결속성 측정 요소를 한글로 전환하는 방법을 제시하였다. 하지만, 이는 한글 텍스트를 영어로 번역해서 결속성을 측정하는 한계를 지닌다. [6]에서는 한국어 텍스트 자체의 결속성을 측정할 수 있는 최초의 도구인 Auto-Kohesion을 제안하였다. Auto-Kohesion은 텍스트 표층 구조, 어휘정보 측정치, 문장 의미 구조, 텍스트 통사구조, 대명사 측정치, 텍스트 연결어 항목을 분석해 이를 수치로 표현하였다. Java를 통해 응용 프로그램을 개발하여 다른 결속성 분석 도구에 비해 활용성 측면에서 뛰어나다는 장점이 있으나, 형태소 태깅이 없어 같은 단어 형태지만 다른 역할을 하는

단어를 구분하지 못하는 결점이 존재한다. 한국어 결속성 측정도구는 대부분 구조적 결속성 측정에 초점을 맞추고 있고, 의미적 결속성은 고려하지 않는다. 또한 구조적 결속성 측정 방법 또한 한국어의 특성을 제대로 반영하지 못하고 있다.

### 2.2 해외의 결속성 자동 측정 연구

영어 결속성 측정 도구로는 대표적으로 TAACO [8]가 있다. TAACO는 영어 문장의 결속성을 측정하기 위해 단어의 태그 비율 분석, 문장(문단) 간 중복 태그 분석, 유의어·동의어 분석, 접속사의 비율을 평가 지표로 사용하는 결속성 측정 도구이다. 하지만 TAACO의 의미적 결속성 측정방식은 LSA와 Word2Vec에 기반하여, BERT[9]를 기반으로 하는 최신 의미적 유사도 측정 방식에 비해 정확도가 떨어진다. [10]은 BERT를 사용하여 의미적 결속성을 측정한 최초의 연구이다. [10]에서는 두 문장 간 상호참조 (Coreference Resolution), 주술관계, 동격 (Bridging) 관계를 측정하여 일본어 결속성을 시각화한 웹페이지를 제시한다. 하지만, [10]에서 사용한 측정 지표를 한국어 결속성 측정에 반영하기 어렵다. 또한, 정확한 결속성 측정을 위해서 구조적 결속성도 함께 제시되어야 한다.

## 3. 제안 모델

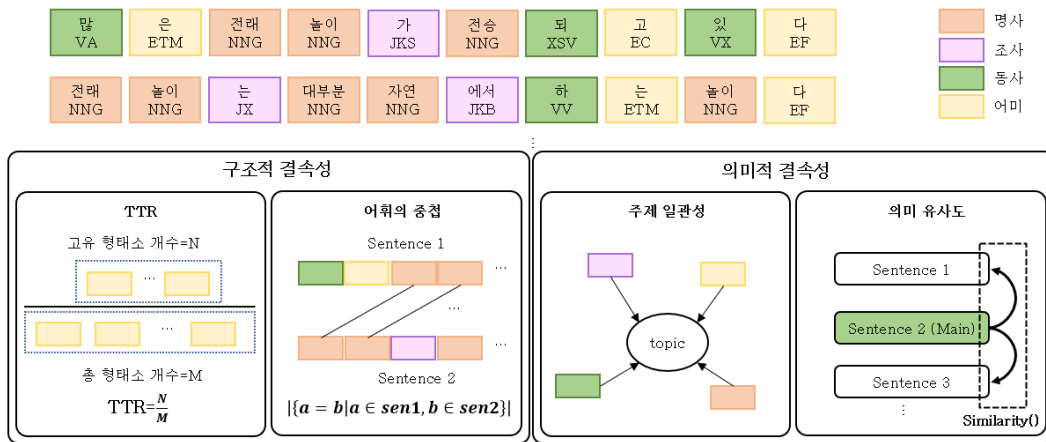
본 논문에서는 한국어 텍스트 도구인 KorCAT을 제안한다. KorCAT은 문장의 구조에 관한 구조적 결속성 측정 방법과 주제 등 의미와 관련된 의미적 결속성 측정 방법으로 나누어 한 문서에 대한 두가지 결속성을 측정하는 방법론이다. 본 논문의 실험을 통해 한국어 텍스트의 의미관계 분석 및 논술형 문항 채점에 채점 근거를 제공하는 도구로서의 가능성을 탐색해보고자 한다.

### 3.1. 구조적 결속성

구조적 결속성 측정 방법은 텍스트가 구조적 형태소를 얼마나 내포하고 있는지 측정한다. 형태소기반의 구조적 결속성 지표 산출 방법은 다음과 같다.

**어휘 다양성 (TTR, Type/Token Ratio):** 총 형태소 수에 대한 서로 다른 고유 형태소의 비율이다. 본 논문에서는 kkmma[11]를 기반으로 토큰화를 수행하였다. 토큰화를 수행하는 형태소의 종류는 어휘 형태소 (Content word), 문법 형태소 (Function word), 명사, 동사, 형용사, 부사, 대명사, bi-gram, tri-gram, 접속사이다. (고유 형태소 수)/(총 형태소 수)

그림 1. 결속성 측정 방안



을 수식으로 하여 형태소 종류별 지표를 측정하였다. 표 1은 어휘 다양성 지표를 통해 문장 내의 형태소 비율을 측정한 예시를 보여준다.

표 1. 어휘 다양성 측정 예시

문장	전래 놀이는 대부분 자연에서 하는 놀이다	
형태소 분석 결과	전래(NNG, 명사) 놀이(NNG, 명사) 는(JX, 보격조사) 대부분(NNG, 명사) 자연(NNG, 명사) 에서(JKB, 부사격조사) 하(VV, 동사) 는(ETM, 관형형전성어미) 놀(NNG, 명사) 이(VCP, 긍정지정사) 다(EF, 종결어미)	
TTR	어휘 형태소	전래, 놀이, 대부분, 자연, 하, 놀이 = 5/6
	문법 형태소	는, 에서, 는, 다 = 3/4
	명사	전래, 놀이, 대부분, 자연, 놀이 = 4/5
	동사	하 = 1/1
	형용사	-
	부사	-
	대명사	-
	Bi-gram	전래 놀이, 놀이는, 는 대부분, 대부분 자연, 자연 에서, 에서 하, 하는, 는 놀이, 놀이다. = 9/9
Tri-gram	전래 놀이는, 놀이는 대부분, 는 대부분 자연, 대부분 자연에서, 자연에서 하, 에서 하는, 하는 놀이, 는 놀이다. = 8/8	

**어휘의 중첩 (Lexical overlap):** 인접한 문장, 문단 간 중복되는 고유 형태소의 비율이다. 고려하는 형태소의 종류는 어휘 형태소, 문법 형태소, 명사, 동사, 형용사, 부사, 대명사, bi-gram, tri-gram이다. 인접한 두 문장을 기준으로 각 문장이 가진 형태소 목록을 추출한 후 이를 대조해 같은 용언과 같은 형태소 태그를 갖고 있는 용언을 파악한다. 표 2는 두 문장 간의 어휘 중첩을 측정한 예시를 보여준다. 문장 2

를 구성하는 형태소 중 문장 1에 나타난 적이 있는 형태소를 중첩된 형태소로 판단해 이를 측정하였다. 문장1과 2의 형태소 분석 결과 중 볼드체 부분은 동일한 형태소와 동일한 용언을 나타내었다. 또한 문장 2의 분석 결과 중 문장 1과 중복되어 나타나는 부분은 밑줄로 표현하였다.

표 2. 어휘의 중첩 측정 예시

문장 1	많은 전래 놀이가 전승되고 있다.
형태소 분석 결과 (문장 1)	많(VA, 형용사), 은(ETM, 관형형전성어미), 전래(NNG, 일반명사), 놀이(NNG, 일반명사), 가(JKS, 주격조사), 전승(NNG, 일반명사), 되(XSV, 동사파생접미사), 고(EC, 연결어미), 있(VX, 보조용언), 다(EF, 종결어미)
문장 2	전래 놀이는 대부분 자연에서 하는 놀이다.
형태소 분석 결과 (문장 2)	전래(NNG, 명사), 놀이(NNG, 명사), 는(JX, 보격조사), 대부분(NNG, 명사), 자연(NNG, 명사), 에서(JKB, 부사격조사), 하(VV, 동사), 는(ETM, 관형형전성어미), 놀이(NNG, 명사), 다(EF, 종결어미)
어휘의 중첩	5

구조적 결속성 측정 방법은 위 2가지를 기초로 하여 측정하였다. 어휘 다양성이 높을수록 다양한 구조적 형태소를 사용하여 텍스트를 작성한 것으로 판단되고, 어휘의 중첩 수치가 높을수록 구조적으로 유사한 형태의 텍스트를 작성한 것으로 판단한다.

### 3.2. 의미적 결속성

의미적 결속성은 주제와 문장 간 유사도를 측정하는 것을 목표로 한다. 방법은 다음과 같다.

**주제의 일관성:** 의미적 결속성이 높은 텍스트는 하나의 주제를 중심으로 내용을 통일성 있게 서술했을 가능성이 높다. 이를 판단하기 위해

Top2Vec [12]을 활용하여 문서의 주제를 추출한다. 문장 임베딩 벡터 값과 문장들을 구성하는 단어의 벡터 값을 하나의 벡터 공간에 배치시키고, 문장 벡터들 사이의 중간 값에 위치한 Topic 을 해당 텍스트의 주제로 선정한다.

**문장 간 의미적 유사도:** 텍스트 내 문장들의 유사도를 판단하기 위해 SBERT 의 STS (Semantic Textual Similarity)를 사용하였다 [9]. 유사도 판단에는 두번의 측정 과정을 거친다. 첫번째로 주제 문장을 파악한다. Top2Vec 을 통해 추출한 주제 단어와 각 문장 사이의 유사도 측정을 통해 가장 주제에 가까운 주제 문장을 산출한다. 이후 주제 문장과 이외의 문장들 사이의 유사도를 비교해 주제 문장에 대해 다른 문장들이 얼마나 집중되게 분포되어 있는지 수치로 환산한다.

표 3. 의미적 결속성 측정 예시

<b>텍스트</b>	옛날 아이들은 윷놀이, 자치기, 팽이치기와 같은 전래 놀이를 하면서 친구들과 즐겁게 놀았다. 대부분의 전래 놀이는 바깥에서 한다. ... 이와 같이 전래 놀이를 하면 좋은 점이 많다. 우리 모두 전래 놀이에 관심을 가지고 즐겨 하자.		
<b>주제</b>	놀이		
<b>문장</b>	<b>주제 유사도</b>	<b>문장 유사도</b>	
옛날 아이들은 윷놀이, 자치기, 팽이치기와 같은 전래 놀이를 하면서 친구들과 즐겁게 놀았다.	0.32	0.46	
<b>(주제 문장)</b> 대부분의 전래 놀이는 바깥에서 한다.	0.38	1.0	
...			
이와 같이 전래 놀이를 하면 좋은 점이 많다.	0.32	0.67	
우리 모두 전래 놀이에 관심을 가지고 즐겨 하자.	0.37	0.65	

표 3 은 의미적 결속성을 측정한 예시를 보여준다. 예시 텍스트에서 주제문장을 선정, 주제문장과 다른 문장들 간의 문장 유사도를 측정한다. 또한, 본 텍스트의 주제인 ‘놀이’와 각 문장들이 얼마나 밀접한 관계가 있는지 주제 유사도를 통해 측정이 가능하다.

### 3.3. 최종 점수 계산

본 논문에서 제시한 방법을 기반으로 평가를 진행하면 텍스트에 대해 어휘 형태소, 문법 형태소, 명사, 동사, 형용사, 부사, 대명사, bi-gram, tri-gram 의 TTR, adjacent overlap 등 44개의 지표가 산출된다. 이 점수는 교사나 평가자에게 제공되어, 학생 글의 언어 간, 내용 간 의미관계가 얼마나 긴밀한지,

그리고 어떠한 요소 측면에서 교육적 개입이 필요한 지 등을 판단하는 척도로 활용될 수 있을 것이다.

## 4. 실험

### 4.1. 실험 세팅

정확한 결속성 측정을 위해 수준별 텍스트를 확보할 필요가 있다. 학생들의 글보다는 전문가의 글을 분석해 제안 방법을 검증하는 것이 신뢰도가 높다고 판단하여, 초, 중, 고등학교 국어 및 국어 관련 교과서의 텍스트를 확보하여 실험을 진행하였다. 의미적 결속성을 측정하는 모델의 훈련을 위한 데이터로 카카오 브레인에서 제공하는 KLUESTS [14] 데이터를 사용하였다. 데이터 전처리에는 형태소 분석 과정이 포함되어 있는데, 이 과정을 위해 상세한 연결어미 단위까지 분석이 가능한 kkmma 를 사용하였다.

### 4.2. 구조적 결속성 측정 및 측정 결과

본 논문에서는 구조적 결속성 측정하기 위해, 우선, 각 문장을 kkmma 를 활용해 형태소 태깅을 실행하고, 각 태그에 알맞은 형태소를 표시한다. 해당 태그들을 기반으로 모든 형태소를 어휘 형태소, 문법 형태소, 명사, 동사, 형용사로 분류해 개수를 측정하고, 중복을 방지하기 위해 형태소의 종류도 측정하는 방식을 택하였다.

표 4. 텍스트 TTR 측정 결과

TTR	초	중	고
어휘형태소	0.743	0.646	0.657
문법형태소	0.374	0.240	0.265
명사	0.555	0.484	0.498
동사	0.760	0.635	0.673
형용사	0.790	0.785	0.730
부사	0.798	0.883	0.790
Bi-gram	0.965	0.961	0.962
Tri-gram	0.991	0.991	0.989
연결어	0.798	0.883	0.790

표 4 는 초, 중, 고 텍스트 80 개의 TTR 비율을 보여준다. 전체 형태소를 제외한 나머지 수치는 고유 형태소의 비율이며 고등학교 교과서의 경우 문법 형태소를 제외하고 고유 형태소의 비율이 높은 편이다. 즉, 다양한 표현을 사용해서 텍스트를 구성했다고 볼 수 있다.

표 5 는 인접 문장간 중복되는 형태소의 개수를 측정한 표이다. 인접의 기준은 i 번째 문장과 i+1 번째 문장으로, 이 둘 사이의 중복되는 형태소의 개수를 측정하여 평균을 산출하였다. 어휘 형태소의 경우는 “~할 것이다.” 등 문장의 끝 부분에서

중복되는 경향이 있으며, 형용사가 가장 적은 중복 비율을 보인다. 표에서 드러나듯이 중학교 교과서의 지수가 초등학교나 고등학교 교과서보다 현저히 낮는데, 여기에는 장르 구성, 텍스트의 난이도, 교육 내용 등 다양한 요인이 관여하였을 것으로 보인다. 향후 국어교육학적 논의를 통해 그 원인을 보다 정밀하게 규명해야 할 것이다.

표 5. 인접 문장간 중복되는 형태소 개수

	어휘 형태소	문법 형태소	명사	동사	형용사	부사
초	13.4	19.4	30	6.64	0.78	0.98
중	35.1	52.6	26.7	18.1	1.94	2.8
고	32.1	47.6	23.4	17.1	1.75	2.65

### 4.3. 의미적 결속성 측정 및 측정 결과

본 논문에서 의미적 결속성은 각 문장이 의미적으로 얼마나 잘 이어져 있고, 하나의 주제에 대해 어느 정도로 응집되어 있는지를 측정한다. 이를 위해 Top2Vec 과 STS 를 사용한다. Top2Vec 은 텍스트 전체 주제를 찾고 응집도 점수를 측정하는데 사용한다. STS 는 주제 문장을 파악하고 타 문장과 주제 문장 간 유사도를 측정하는데 사용한다.

텍스트 전체 주제는 Top2Vec 으로 문장에 나타난 단어들을 하나의 벡터 공간에 위치시킨 후 HDBSCAN[13]을 기반으로 클러스터링 하여 산출한다. 하지만 Top2Vec 을 한국어에 적용하기 위해서는 불용어 제거와 같은 전처리 작업이 필요하다. 추가적으로 한국어 기반 주제 분류 (Topic Classification) 데이터를 모델에 학습시키는 과정을 진행하여 주제와 주제 문장을 추출하였다. 이후, 주제 문장과 나머지 문장을 코사인 유사도로 비교하여 평균값을 산출하였다. 표 6 은 초, 중, 고 교과서 텍스트의 의미적 결속성 측정결과를 보여준다.

표 6. 텍스트 의미적 결속성 측정

	주제 결속성	주제 유사도
초	0.3135	0.4184
중	0.3321	0.4237
고	0.3346	0.4335

표 6 은 주제 결속성과 주제 유사도를 측정한 결과를 보여준다. 두 항목은 0~1 사이의 값으로 측정되며, 초, 중, 고 교과서 텍스트에서 모두 비슷한 수치가 측정됨을 알 수 있다. 교과서는 전문가가 작성한 텍스트로 이루어져 있기 때문에, 텍스트의 수준이

다르다고 하더라도 주제 결속성이나 주제 유사도 수치가 비교적 높은 수준으로 비슷하게 유지됨을 보여주는 결과이다. KorCAT 으로 측정된 전문가 수준의 의미적 결속성 측정값을 기반으로 학생이나 일반인이 작성한 텍스트의 의미적 결속성을 측정하면, 텍스트의 작문 수준을 평가하는 데 도움을 주는 지표로 활용될 수 있을 것이다.

### 4.4 최종 분석 결과

KorCAT 을 통해 측정한 결과를 분석해보면 텍스트 수준에 따라 구조적 결속성이 다른 것을 확인할 수 있다. 초, 중등 교과서에 비해 고등 교과서에 좀 더 다양한 연결어, 어휘형태소, 명사, 부사 등이 사용되었다. 이를 통해 학년이 증가함에 따라 학습자가 다양한 결속기제 관련 어휘를 접하게 된다는 결론을 도출할 수 있었다.

하지만, 의미적 결속성은 문장 간의 유사함을 측정하는 척도이기 때문에, 텍스트의 수준과 큰 상관관계가 없음을 확인할 수 있다. 의미적 결속성은 텍스트의 수준이 아니라 텍스트가 주제에 맞게 일관적인 문장으로 구성이 되어있는지 의미적으로 유사한 문장으로 이루어져 있는지를 측정하는 척도이기 때문에, 교과서 텍스트와 같이 전문가 수준에서 작성한 텍스트라면 텍스트의 수준과 관계없이 주제의 일관성을 유지하면서 동시에 의미가 유사한 문장으로 양질의 텍스트가 이루어져 있음을 알 수 있다. KorCAT 을 통해 측정된 전문가 수준의 텍스트 결속성 측정값을 기준으로 하고, 학생 또는 일반인이 작성한 텍스트를 수집하여 결속성을 비교하는 실험을 추가적으로 진행한다면, KorCAT 을 통해 대규모 작문 평가를 위한 결속성 측정도구로 충분히 활용될 수 있음을 확인할 수 있을 것이다.

### 5. 결론

본 논문에서는 한국어 텍스트의 구조적, 의미적 결속성을 평가할 수 있는 도구인 KorCAT을 제안하였다. KorCAT을 통해 제안한 한국어 텍스트 결속성 지표를 검증하기 위해 초, 중, 고등학교 교과서 텍스트를 확보해 전문가가 작성한 수준별 텍스트들의 결속성을 측정해 보았다.

향후 연구로는 본 논문에서 측정한 전문가 텍스트의 결속성 지표를 기준으로 학생 및 일반인이 작성한 텍스트의 결속성 평가를 진행해보고자 한다. 또한, 실제 국문학 및 국어교육학의 전문가들에게 자문을 요청하여 KorCAT의 정확성과 필요성을 추가로 검증하려 한다.

**감사의 글**

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2022-0-00448, 인간처럼 회상이 가능한 인공 신경망 지속 학습 플랫폼 개발, No.RS-2022-00155915, 인공지능융합혁신인재양성(인하대학교))과 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2022R1C1C1012408, 우수신진연구).

**참고문헌**

[1] 길호현, "텍스트의 응집성(coherence) 형성 요소와 양상에 대한 연구," 독서 연구, Vol.0 No.56, 193~224p, 2020

[2] 정희모, "한국어 교육과 결속성(cohesion) 및 응집성(coherence)의문제", 리터러시 연구, Vol.10 No.4, 89~123p, 2019

[3] 이재승, "읽기와 쓰기 행위에서 결속 구조의 의미와 지도", 국어교육 110권 0호, 91~111P, 2003

[4] 손마정, "한국어 교육에서의 텍스트 결속 표지 연구", 우리말학회 학술대회지, 7~15p, 2018

[5] 권민지, "[교육칼럼]2028년부터 미래형 수능 도입", 조선에듀, 2021년 7월 29일자

[6] 전문기, "한국어 텍스트 분석 시스템의 개발과 활용," *중견연구자지원사업*, (2017)

[7] 김영신, "영한번역에서 결속장치(Cohesive Ties) 전환연구 : 접속사(Conjunction)를 중심으로," *통번역교육연구*, Vol.7, No.1, pp 99-115. 2009.

[8] S. A. Crossley, K. Kyle, and D. S. McNamara, "The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion," *Behavior research method*, Vol. 48, No. 4, pp. 1227-1237, 2016.

[9] Jacob Devlin and Ming-Wei Chang, "BERT-Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Vol.1, pp 4171-4186. 2019.

[10] Nobuhiro Ueda and Daisuke Kawahara, "BERT-based Cohesion Analysis of Japanese Texts," *Proceedings of the 28<sup>th</sup> International Conference on Computational Linguistics(COLING)*, Vol.1, pp 1323-1333. 2020.

[11] 이동주, 연종흠, "꼬꼬마 : 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구," *정보과학회 논문지: 컴퓨팅의 실제 및 레터*, Vol.16, No.11, pp 1046-1050. 2010.

[12] D. Angelo, "Top2Vec:Distributed Representations of Topics," *arXiv preprint arXiv:2008.09470*, 2020.

[13] L. McInnes and J. Healy, "Accelerated Hierarchical Density Based Clustering," *Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp 33-42. 2017.

[14] Sungjoon Park, "KLUE:Korean Language Understanding Evaluation," *arXiv preprint arXiv:2105.09680v4*, 2021