자연언어처리를 활용한 텍스트 연구 분야의 비교를 통한 자동채점 변인 탐색*

이현준** · 박영민***

차 례

Ⅰ. 작문 평가의 쟁점

작문 평가는 평가의 관점이나 목적을 어디에 두는가에 따라 평가의 과제, 대상, 기준, 방식, 피드백 등이 달라진다. 기본적으로 교육 평가는 학생이 가진 능력을 진단하고 그에 따른 결과를 교육적으로 해석하여 피드백을 제공하는 것을 목적으로 한다. 박영민 외(2016)에서 작문 평가를 “글을 쓰는 데에 필요한 능력을 다양한 측면에서 측정하고, 이들 결과에 근거하여 학생의 작문 능력으로 교육적으로 해석하고 판단하여 학생에게 적절한 피드백을 제공하는 행위”라고 정의하였다. 따라서 작문 평가는

* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF – S1A5A2A01027380)
** 1저자, 중원대학교 초빙교원
*** 교신저자, 한국교원대학교 국어교육과 교수
DOI : 10.31565/korrow.2019..41.009

학생의 현재 작문 능력을 정확히 평가해야 한다. 정확한 평가의 의미는 온전한 작문 능력을 평가할 수 있어야 함을 의미한다. 평가자는 평가를 위해 어떤 요소를, 어떠한 방식으로, 어느 정도의 기준점으로 평가 결과를 도출할 것인지에 대해 고민하여 기준의 일관된 적용과 정확한 판단을 내릴 수 있어야 한다.

그러나 작문 평가는 평가에 소요되는 시간과 노력의 양, 비용의 문제 등으로 인해 현실적인 제약에 부딪히곤 한다. 이러한 제약을 극복하고 실제 글을 평가한다고 하더라도 평가 결과에 대한 정확성과 일관성의 문제에서 자유로울 수 없다. 인간 평가자는 평가 과정에서 글이 지니고 있는 내용이나 의미, 정보의 양, 차별성이나 독창성과 같은 글의 가치를 수량화, 정량화하여 측정하거나 판정을 내리지 않는다. 이는 작문 평가를 위해 하나의 평가 기준을 적용하더라도 평가자마다 평가에 활용하는 정보나 단서가 다를 수 있음을 의미한다. 무엇을 좋은 글 또는 잘 쓴 글이라고 판정할 것인지에 대한 합의에서부터 어느 정도의 수준 이상을 상위 등급으로 판정할 것인지는 평가자, 평가 과제, 평가 기준, 평가 대상 등에 따라 달라진다.

또한 동일한 점수나 등급으로 판정을 받은 글이 모두 유사한 특성을 지니고 있지 않다는 점도 평가 결과를 해석하고 학생들에게 유의미한 정보를 제시함에 있어 한계를 지닌다. 평가 결과의 차이가 글의 어떠한 특성에서 기인하는지 그리고 그 특성의 어느 정도의 차이에 의한 것인지를 설명함에 있어 어려움이 따른다. 예를 들어 학교 현장에서 학생이 자기가 왜 이러한 평가 결과를 얻었는지에 대한 피드백을 요구할 경우 평가자는 명확한 답변을 제공하기에 어려움이 있다. 더 좋은 평가를 받은 글과 비교, 대조하여 설명한다고 하더라도 글의 어떤 부분이 어떻게 다른지 평가 결과에 수긍하고 납득할 만한 정보를 제시하기 위한 고민이 요구된다. 또한 좋은 평가를 받는 글이 꼭 하나의 기준만 있는 것도 아니다.

작문 평가는 필자의 작문 과정과 작문 결과물인 텍스트의 특성에 의해 평가의 어려움이 더욱 가중된다. 학교에서는 학생들에게 가르친 것, 그리고 좋은 글을 쓸 수 있는 방법으로 내용, 조직, 표현 등을 가르치지만 세부 내용을 평가하기 위해서는 결국 완성된 글의 형태가 필요하다. 조직하기에 대한 단원을 학습하고 조직을 잘 할 수 있는지에 대한 평가를 하는 과정에서 글을 작성해야하는데 조직은 내용이나 표현에서 자유로울 수 없다. Smith & Dunstan(1998)에서 글이 아닌 학습 결과에 대한 평가가 이루어져야 함을 주장하였지만 글의 특정 부분만을 분류하여 인식하고 독립적으로 평가를 내리기에는 읽기라는 인지 과정이 이루어지는 과정을 살펴볼 때 역시 난해한 부분이 존재할 수밖에 없다.

이처럼 작문 평가는 작문 자체가 지니고 있는 특징, 그리고 작문 과제, 평가자, 평가 기준, 평가 순서 등을 포함한 평가 맥락에 따른 특징에서 기인하는 어려움이 따른다. 이를 정리하자면 평가의 배타성 또는 독립성 그리고 평가의 보편성, 정확성, 일관성, 마지막으로 평가의 정보성 측면에서 인간 평가에 의한 현재의 방식에 논란과 난점이 많다. 이를 위해 1960년대부터 Page는 에세이 자동 채점(AEE)에 관심을 가졌으며 글에서 관심 있는 특징을 대략적으로나마 수량화하여 평가에 활용하기 위해 글을 구성하고 있는 요소들의 기본 척도를 통해 글의 특징을 설명하고 평가를 내리고자 했다. 그 후 자연언어처리(NLP)의 발전으로 자동으로 글의 특징을 추출하고 글의 수준을 평가할 수 있는 수준까지 발전했다.

이제 인간 평가의 방식에서 자동 채점의 방식으로 일부 또는 전체 과정을 대체할 수 있는 여건이 갖추어진 상황에서 작문 교육, 특히 작문 평가에서 자동 채점을 위해 필요한 기본 여건에 대한 논의가 필요하다. 이 연구에서는 자연언어처리를 통한 기존의 텍스트 연구 분야에 대한 이해와 비교를 통해 자동 채점을 위해 작문 교육 분야에서 이루어져야 할 토대 연구가 무엇인지에 대해 탐색하고자 한다.

Ⅱ. 인간 평가자의 작문 평가

1. 작문 평가를 위한 기본 전제

독자의 읽기 과정에 대한 연구는 아직 명확히 밝혀지지 않은 측면이 많다. 더욱이 독자가 읽고 의미를 이해한 후 그것에 대한 판정을 내려야 하는 작문 평가를 위한 읽기는 이해를 위한 읽기보다 더 복잡하며 개입되는 영향요인이 더 많은 것이다(Golding et al., 2014). 인간 평가자의 평가 방식에 대한 이해는 기본적으로 인간의 읽기 과정에 대한 이해를 통해 이루어진다. 읽기라는 현상을 해석하기 위한 여러 이론이 있지만 이 연구에서는 기계식 평가와의 비교를 위해 인지주의에 기반한 인간의 읽기 이해 과정을 살펴봄으로써 작문 평가에서 인간 평가자의 평가 방식을 살펴보고자 한다. 결과 중심 작문 평가는 학생들의 작문 과정을 평가하는 것이 아니라, 산출 결과로서 생산된 글을 읽고 평가하는 형태이다. 이때 작문 평가자의 글 읽기 과정은 글에 대한 가치 부여 또는 평가 과정의 시작점이며, 평가 결과를 내리고 피드백 정보를 구성하기까지 수많은 정보처리 과정을 거치게 된다(Pressley, Borkwski, & Schneider, 1989). 작문 평가는 피평가자의 작문 수행 능력이 어떠한가에 대한 기본적인 정보 처리에서부터, 필요에 따라 앞으로 더 개선하고 보완해야 할 능력이 어떤 부분이며 작문 수행에서의 장단점 분석 등을 제시해 줄 필요성이 있다. 이를 위해서는 평가 기준과 평가 준거, 그리고 평가 규칙 등이 필요하다.

이때 평가 준거는 평가의 기준, 또는 판단의 참조점을 의미한다. 작문 평가에서의 평가 준거는 성취 기준을 어느 정도 수준 이상으로 달성한

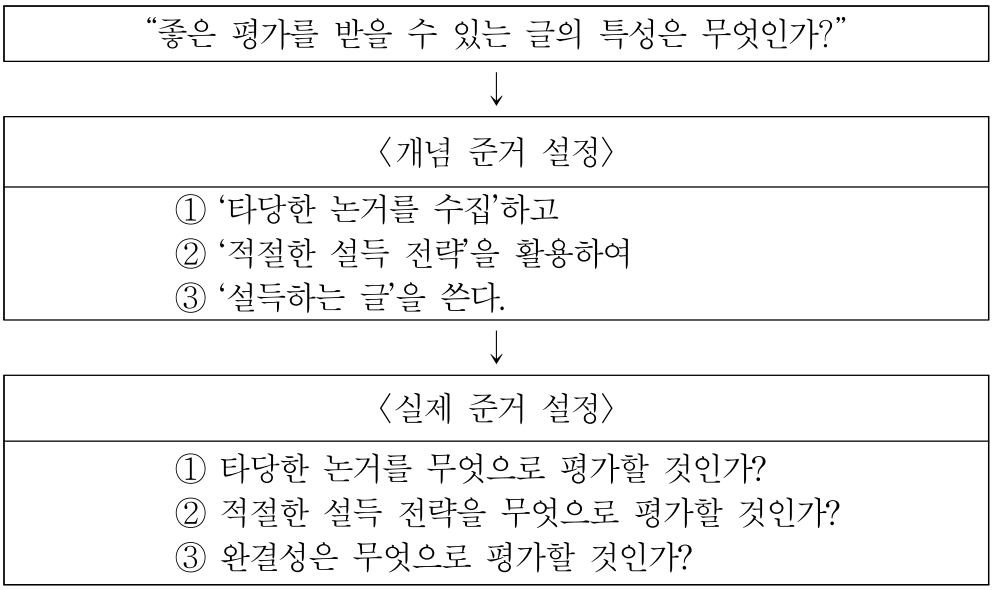
‘잘 쓴 글’을 판단하는 기준이 된다. 예를 들어 ‘타당한 논거를 수집하고 적절한 설득 전략을 활용하여 설득하는 글을 쓴다.’라는 성취 기준이 있을 때 평가를 위해서는 평가 준거와 평가 기준을 마련해야 한다. 평가자는 학생들이 해당 성취기준을 성공적으로 달성하기 위한 작문 능력을 갖추고 있는지를 평가하고 판단하고자 하는 목적에 있다.

무엇을 성공적인 작문 능력으로 볼 것인지를 측정하기 위해서는 개념 준거를 먼저 설정해야 한다. 앞서 제시한 성취 기준에서는 성공적인 작문의 기준으로 타당한 논거를 수립할 수 있어야 하며, 논거를 적절한 설득 전략을 활용하여 한 편의 완성된 글의 형태로 구성할 수 있는 능력으로 제시하고 있다. 그러나 성취 기준이나 평가 준비 단계에서는 준거가 개념 준거 수준의 상태인 경우가 많다. 개념 준거는 실질적으로 측정할 수 없는 추상적이고 이론적인 개념 상태를 의미한다. 타당한 논거 수집, 적절한 설득 전략, 완결성 있는 한 편의 설득하는 글 등은 측정하고 판정하기 어려운 상태에 있다. 무엇을 타당한 논거 수집의 결과로 볼 것인지, 적절한 설득 전략으로 볼 것인지, 그리고 완결성을 갖춘 한 편의 설득하는 글을 무엇으로 평가할 것인지에 대한 준거 설정이 필요하다.

이러한 개념 준거를 실제로 측정하기 위한 현실적인 요인으로 바꾸는 것이 바로 ‘실제 준거’이다. 실제 준거는 평가자가 측정하거나 평가하는 데 사용하는 조작적 혹은 실제적인 기준을 의미한다. 타당한 논거 수집을 무엇으로 구체화하여 측정할 것인가부터 한 편의 완결된 설득하는 글의 형태를 무엇으로 측정할 것인지를 설정하는 것이다. 개념 준거와 실제 준거 간의 관계는 하나 이상의 실제 준거를 선택하여 개념 준거에 대해 정밀하고 정확한 추정치를 얻을 수 있어야 한다. 즉 평가 준거는 ‘무엇’을 평가할 것인가를 규명하며 평가 활동의 내용과 범위, 또는 방향을 제시해주는 역할을 한다(Muchinsky & Culbertson; 유태용 역, 2016).

위의 예시 성취 기준은 이미 개념 준거는 갖추고 있다고 볼 수 있다. 작문 교육의 목적은 학생들이 ‘유능한 필자’가 되도록 하는 것에 목적을

준다. 유능한 필자가 무엇인지 그리고 유능한 필자는 무엇을 할 수 있어야 하는지에 대한 고민의 결과를 각 성취 기준인 개념 준거로 설정해 둔 것이다. 이에 따라 유능한 필자에 대한 하나의 정의를 '타당한 논거를 수집하고 적절한 설득 전략을 활용하여 설득하는 글을 쓸' 수 있는 상태로 규정하였다. 이처럼 성취 기준은 이미 하나의 개념 준거의 역할을 해준다. 그러나 개념 준거가 실제 준거의 역할이 되어 주지는 못한다. 개념 준거를 측정하기 위한 실제적인 기준이 마련되어 있지 않기 때문이다. 따라서 각 개념 준거를 측정하고 판단하기 위한 구체적 실제 준거가 마련되어야 한다. 이때 개념 준거와 실제 준거가 일치하거나 유사한 정도를 준거적절성이라고 한다.



평가 기준은 평가 준거에 대해 목표 달성 수준, 기대하는 변화의 정도, 의도하는 특성이 나타나는 정도, 변화 또는 성취를 나타내는 점수, 의도한 장점이 표현되는 맥락이나 상황의 수준, 원하는 가치가 산출되는 정도나 범위 등을 의미한다(장성연, 2017: 133). 작문 평가라면 부여된 작문 상황에서 얻어낸 특정 점수나 성취의 구체적인 정도나 수준을 의미한다. 평가 기준은 각 준거에 대한 평가 자료를 중심으로 그 준거의 속성

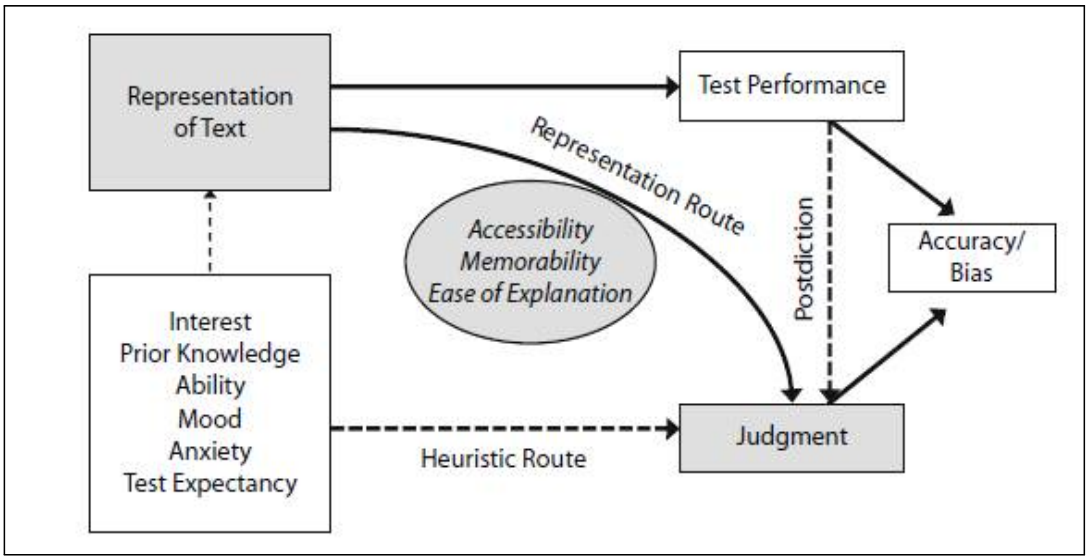
과 내용, 또는 그로 인한 산출 및 결과에 관하여 어느 정도/얼마만큼이 바람직한 수준, 점수, 기준점이 되는가를 나타내는 것이다. 평가 준거마다 하나 이상의 평가 지표가 수집되고 그에 따라 평가 기준이 설정되는데 평가 기준은 평가 자료를 통해 표현되기 때문에 어떤 평가 자료를 갖추느냐에 따라 기준은 달라진다. 평가 기준은 그 자체로서는 특정한 의미를 지니고 있지 않다. 평가 준거 및 평가 지표와의 관계에 의해서만 그 의미가 부여된다. 평가 기준을 마련하기 위해서는 이미 설정된 평가 준거에 근거해야 한다(구경호, 2015). 어떤 준거에 대한 기준을 어떤 자료와 지표를 사용하여 설정한 것인지를 고려해야 하는 것이다. 하나의 평가 준거에 하나 또는 여러 평가 기준이 있을 수 있으며 여러 종류의 평가 준거에 대해 하나의 평가 기준만을 설정하여 활용할 수도 있다. 즉 평가 기준은 '평가 준거의' 또는 '평가 준거에 근거하여' 바람직한 수준이나 정도, 상태 또는 질 등을 나타낸다.

2. 인간 평가자의 작문 평가 방식

작문 평가, 특히 직접 평가에서는 작성된 결과물인 글을 평가하는 것이다. 이때 평가의 기준은 작문 교육과정 또는 작문 교과목에서 성취기준으로 다루어지며 작문 능력 향상을 위해 교수-학습으로 다루어진 내용을 기준으로 한다. 주로 작문 평가의 준거는 '내용, 조직, 표현(어조 및 태도), 단어 선택, 형식 및 어법'의 다섯 가지로 이루어진다. 이 준거들은 능숙한 필자의 결과물을 구체화하고 세분화할 수 있는 역할을 한다. 평가를 위해 평가 준거와 평가 기준을 마련했다고 하더라도 작문 평가의 상황에서는 이것을 정확하고 동일하게 적용하는 과정에는 또 다른 문제가 발생한다. 동일한 평가 준거를 활용한다고 하더라도 실제 평가자들이 평가 준거를 어떻게 해석하고 받아들여서 이를 실제 평가 과정에 반영하고 처리하는 방식은 평가자마다 차이를 보인다.

작문 평가에서는 평가 기준에 적합한 정보를 객관적 형태로 얻어낼 수 없으며 따라서 주관적 정보 선택과 판단이 이루어지게 된다. 인간 평가자는 평가를 위해 글을 읽은 후 텍스트 이미지화의 과정을 거친다(Crisp, 2012 ; Freedman & Calfee, 1983). 평가 맥락과 준거를 고려한 평가 과정에서 구현한 텍스트의 이미지에 부합하는 적절한 평가 결과를 정리하는 과정을 거쳐야 하는 것이다(Baird, 2000). 이러한 텍스트 이미지화의 과정에서는 평가자의 인지적 처리 과정에서 '비교'와 '판단'이라는 사고 과정이 매우 중요하다. 인간 평가자는 글에 대한 전체적인 텍스트의 이미지를 통해 평가를 하기도 하고, 글의 부분의 특성들을 이미지화하여 평가를 하기도 한다. 그리고 이러한 글 전체와 부분 간의 특성을 종합하여 평가를 하기도 하는 등 복잡한 사고 처리를 통해 평가 결과를 산출한다(오세영, 2014: 117).

평가자들은 텍스트 이미지들 사이에 비교를 하게 되고 테스트 이미지에 부여된 점수를 비교하여 적절한 평가 결과를 제시하기 위한 정보 처리 과정을 거치게 된다. 그리고 이러한 과정에서 기준이 되는 것이 바로 평가 준거인데, Eckes(2012)는 준거는 텍스트를 이미지화하는 인지적 처리 과정에서 상세한 항목 그대로 기억되어 있기 보다는 덩어리 형태로 텍스트 이미지화의 복잡한 사고 과정너머로 사라지기도 한다고 설명하였다(오세영, 2014: 118). 즉 작문 평가 상황에서 인간 평가자는 실제 준거에 대한 판단을 위해 평가 기준을 활용하지만 평가자마다 활용하는 단서도 다르며 각 단서가 정확히 수치로 환산되는 것이 아닌 하나의 이미지로 저장되고 비교되기 때문에 주관적 판단이 개입할 수밖에 없다. 즉 텍스트의 이미지화 과정에서 이루어지는 '비교'와 '판단'의 기준이 모두 다를 수 있으며 이에 따라 평가 결과 역시 다른 결과를 나타내게 된다. 따라서 각 평가자마다 무엇을 근거로 어떻게 '비교'와 '판단'의 처리 과정을 진행해 가는지가 모두 다르고 명확하지 않다는 점이 작문 평가 과정을 이해함에 있어 어려움을 초래하게 된다.



〔그림 1〕 평가자의 판단을 위한 잠재적 절차
(Griffin, Jee, & Wiley, 2009: 1002)

작문 평가 상황에서 평가자에게 동일한 평가 기준표를 제공한다고 하더라도 평가자마다 학생글을 읽어 가는 읽기 과정에서의 정보 처리 과정의 차이, 그리고 이를 비교하고 판단하기 위한 이미지화의 정보 처리 과정에서 활용하는 정보와 판단의 기준점이 모두 다르다는 점은 인간 작문 평가에 대한 단일한 정의를 내리거나 규칙화하기 어렵다는 것을 의미한다. 그러나 인간 작문 평가자들은 평가를 위해 동일한 절차는 밟아 나간다. 즉 작문평가 준거를 고려하여 텍스트를 읽고 이미지를 형성하는 단계와 각 이미지에 대한 '비교'와 '판단'의 단계를 거친다는 것은 동일하다. 각 이미지를 형성하기 위해 무엇을 활용하는지는 평가자마다 다르지만 수량화하기 어려운 텍스트 정보를 나름대로 수집하여 하나의 이미지로 생성한다는 것, 그리고 각 이미지들의 전체 또는 부분을 비교하여 평가 결과를 산출하는 비교와 판단의 과정은 인간 작문 평가자의 동일한 평가 절차로 확인할 수 있다.

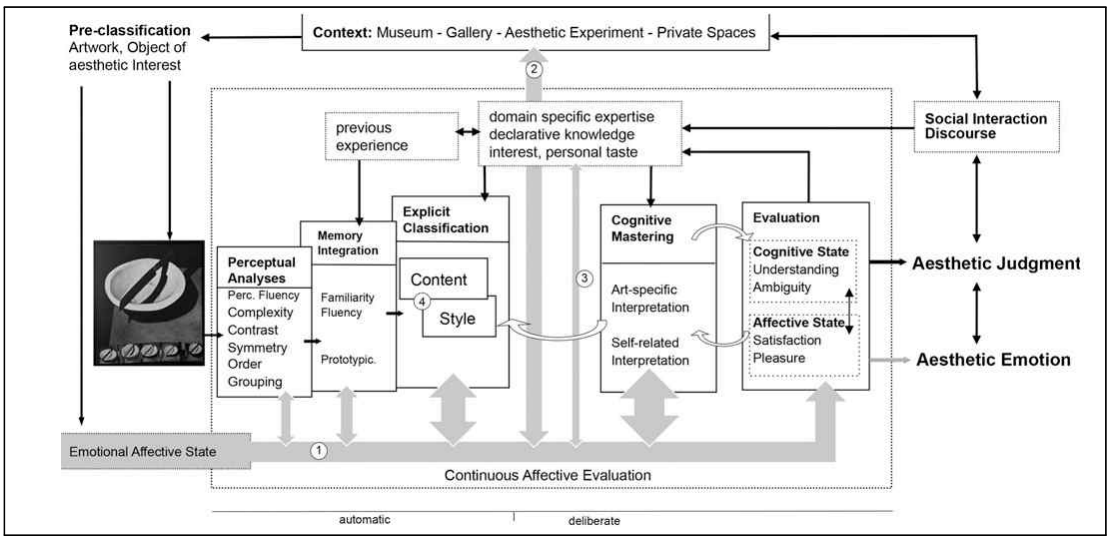
3. 비정형 데이터의 평가 사례

비정형 데이터의 속성을 지닌 대상을 평가하는 인간의 인지 과정을 분석하기 위한 노력은 작문 평가 이외에도 시각 예술 분야인 미술 감상에서도 그 사례를 찾을 수 있다. 시각 예술분야에서 추구하는 '아름다움'에 대해 인간이 작품을 보고 느끼는 과정을 인지 신경 미학으로 접근한다. 인간의 작문 평가와 마찬가지로 미학적 판단 역시 관찰자의 외부 세계에 대한 단순 지각이 아니라 능동적 해석과 재창조의 과정으로 이해하는 것이다.

Kawabata & Zeki(2004)는 실험 참가자들에게 정물화, 인물화, 풍경화, 추상화를 2초 간격으로 300점의 작품을 제시하였다. 짧은 시간 동안 작품을 보고 아름다움의 정도를 1~9점으로 평가하게 하였고 이를 바탕으로 아름다움, 보통, 흉함으로 분류하였다. 참가자들이 작품을 감상할 때 활성화되는 뇌의 부분을 살펴본 결과 아름다움으로 분류된 작품을 감상할 때에는 내측 안와전두엽이 활성되며, 흉함으로 분류된 작품에서는 해당 부위가 크게 저하되는 것을 발견하였다. 미술 작품을 감상할 때 활성화되는 내측 안와전두엽 그리고 전전측 대상회는 보상체계와 관련된 부분으로 인간의 쾌락과 밀접한 관련성을 지니고 있다. 이에 따라 인간의 미학적 판단은 단순한 느낌이 아니라 특정 뇌의 부분이 활성화되는 인지적 판단의 결과임을 확인하였다. 이는 인간의 미학적 판단을 유발하는 특정 원리가 내재되어 있음을 확인한 것이다. 작품이 지니고 있는 특정 요소들이 인지적 쾌감을 유발하며 이를 통해 아름다움이라는 주관적 의사 결정을 내리게 된다.

Leder et al.,(2004)은 인간이 작품을 감상할 때 겪게 되는 절차를 '미적 경험 모델'을 통해 설명하였다. 미적 경험 과정에는 대상에 대한 지각적 분석과 과거의 경험을 통한 암묵적 기억이 통합된다. 그리고 이를 통

해 내용과 형식에 대한 명시적 분류가 이루어지며 예술적 · 자의적 해석의 과정을 통해 대상에 대한 미적 평가가 이루어진다. 이러한 과정에는 작품을 감상하는 맥락이 영향을 미치기도 한다. 이때 지각적 분석, 암묵적 기억의 통합, 명시적 분류는 자동적 과정이며 인지적 통달과 평가는 의도적 과정으로 분류된다. 그러나 명시적 분류 과정에 속하는 내용과 형식이 구체적으로 어떤 항목들을 인지하는지에 대해서는 관찰자마다 다르며 이에 대한 일반적 준거를 마련하지는 못하였다.



〔그림 2〕 미적 경험 모델(Leder et al., 2004: 492)

시각 예술 작품에 대한 신경 미학 관점에서의 주관적 판단은 텍스트를 읽고 '잘 쓴' 또는 '좋은' 글을 평가하기 위한 평가를 위한 읽기의 과정과 유사한 측면을 지니고 있다. 앞서 인간 평가자의 평가 과정에서도 텍스트를 이미지화 하는 단계가 포함되어 있던 것처럼 미적 경험의 과정에서도 내용과 형식에 대한 명시적 분류 작업이 이루어진다. 대상에 대한 아름다움 또는 좋고 나쁨의 미적 비교와 판단 결과의 원인을 파악하기 위해서는 평가를 위한 요소가 무엇인지 확인할 수 있어야 한다. 이는 무엇을 좋은 텍스트의 특징으로 볼 것인지에 대한 정의에서부터 시작된다.

Elgammal et al.,(2017)은 이러한 원리를 바탕으로 인공지능을 통한 미술 작품의 평가를 수행했다. 인공지능을 통한 작품 평가를 위해 인간의 지각과 인식, 그리고 평가 과정과 유사한 지적 능력을 갖춘 알고리즘을 구축하였다. 이 연구에서는 특히 미술 작품의 창의적인 수준을 측정하고자 하였다. 이를 위한 전제 조건은 창의성을 어떻게 측정할 것인가, 즉 창의성을 측정하기 위해 어떤 요소를 수량화할 것인가에 핵심이 있다. 창의성 판단 알고리즘 개발을 위해 활용한 변수로는 작품의 색, 질감, 관점, 그리고 대상을 고려할 수 있도록 하였고 이를 바탕으로 한 매개변수로 창의성을 다루기 위해 독창성과 후세 작품의 영향력을 기준으로 하였다. 이처럼 어떤 대상을 평가하기 위해서는 그 대상의 무엇을 평가할 것인지를 명확히 규정해야 하며(개념 준거) 평가하고자 하는 범주를 산정했다면 해당 평가 범주를 대표할 수 있는 변인(실제 준거)을 탐색해야 한다.

인간이 아닌 컴퓨터를 활용하여 데이터를 분석하고 분류하기 위해 알고리즘이나 모델을 구축할 때에는 대상이 지니고 있는 속성을 최대한 많이 포함하면서도 실제 대상보다는 간단해야 한다. 즉 대상의 속성을 가장 정확하게 설명할 수 있는 변인을 중복과 누락 없이 포함해야 하며 불필요한 속성이나 중복되는 속성은 배제해야 한다. 이때 대상을 설명하기 위해 반드시 포함되어야 하는 정보를 신호(signal)라고 하며 불필요한 정보를 잡음 또는 소음(noise)이라고 한다. 평가를 위해 개념 준거를 수립했다 하더라도 개념 준거의 내용 중에는 실제 준거에 의해 측정할 수 없는 부분이 존재할 수 있다. 이를 준거결핍이라고 한다. 개념 준거를 완벽하게 설명해 낼 수 있는 신호를 찾는다면 가장 좋은 평가준거가 되겠지만 이 준거결핍을 줄일 수 있는 최적의 신호를 찾는 것이 필요하다. 반대로 불필요한 정보, 즉 잡음이나 소음으로 인해 실제 준거가 개념 준거와 관련성이 없는 경우를 준거오염이라고 한다. 준거오염은 실제 준거가 개념 준거가 아닌 다른 것을 측정할 수도 있으며(편파, Bias) 어떠한 것과도

관련되어 있지 않을 수 있다(오류, Error). 즉 정확한 평가를 위해서는 개념 준거와 실제 준거의 교집합을 최대화하고 결핍이나 오류를 최소화하는 것이 중요하다.

작문 평가에 적용할 수 있는 자연언어처리를 통한 텍스트 분석 연구는 기본적으로 각 텍스트에 포함된 단어를 공간으로 벡터화한 벡터공간모델(Vector Space Model)을 기반으로 한다. 각 단어가 하나의 공간을 차지하며 단어의 출현 여부 또는 출현 빈도를 계산한다. 비정형 데이터인 텍스트를 정형화된 형식으로 구현하기 위해서는 텍스트의 기본 구성 단위인 단어 또는 형태소를 기본 분석 단위로 설정하는 것이 가장 일반적인 방식이다. 이후 텍스트를 구성하는 단어의 수가 많기 때문에 차원을 축소하는 방법을 적용하거나 의미적 요소를 고려하기 위해 동시 출현 단어(연어, Co-occurrence)나 잠재 의미 분석(LSA)을 활용하기도 한다.

이렇게 구축된 텍스트 데이터는 번역, 문서 검색과 정보 분석, 문서의 분류, 문서 작성과 요약, 언어 인식, 질의응답 시스템, 맞춤법 검사, 이독성 연구, 문서 유사도 비교, 특정 필자의 고유성(변별성) 판독, 인공 지능 스피커, 감성 분석, 광학 글자 판독(OCR) 등에서 널리 활용되고 있다. 어떤 활용분야이건 간에 텍스트 분석의 방법과 절차에 있어 유사점을 공유하고 있다. 즉 텍스트 분석 척도를 통해 수량화하고, 이렇게 수량화된 정보를 비교, 분석, 분류, 평가하는 데에 활용한다. 활용 분야는 다양하지만 단어를 기본 분석 단위로 하여 단어가 지닌 형태적 특성 또는 의미적 특성을 사용하거나 단어 특성의 조합을 달리하여 적용한다.

텍스트에 대한 자동 채점 역시 '좋은' 또는 '잘 쓴' 글을 무엇으로 정의할 것인지에 대한 합의와 함께 텍스트의 어떤 요소를 텍스트의 특성을 설명할 수 있는 신호로 포착하여 다룰 것이며 어떤 특성을 불필요한 잡음으로 판단하여 평가 과정에서 배제할 것인지에 대한 탐구가 필요하다. 이를 위해 자연언어처리 방식을 활용하여 텍스트 분석 또는 분류 연구에 적용되고 있는 텍스트의 분석 요소와 각 요소를 다루는 방식을 살펴보고

분야별 차이점을 도출하고자 한다. 이를 통해 기존의 작문 평가에서 나타난 문제점을 보완하면서도 평가로서의 가치를 지닐 수 있는 방법을 모색하고자 한다.

Ⅲ. 자연언어처리를 활용한 텍스트 연구 분야

1. 이독성 연구

이독성은 문자 언어에 있어서 글의 내용을 읽고 이해할 수 있는 정도, 즉 텍스트 중심 읽기 자료(reading material)의 쉽고 어려운 정도를 의미한다. 독자가 지닌 내적 요인들과 상황 맥락으로 인한 주관적이고 상대적인 어려움의 정도가 아닌 문장이나 글 자체 특성에서 기인하는 이해하기 쉽고 어려운 정도를 판별하는 것이다. 그러나 최근에는 텍스트 외부 요인인 독자나 질적 요인(의미나 맥락)까지 포함한다. 이독성의 궁극적인 목표는 독자와 텍스트 상호 간의 간계 속에서 어떤 글이 더 어려운 글인지를 판정해 내는 것이다. 대부분의 이독성 관련 연구는 텍스트를 측정할 수 있는 요소로 분절하고 변환하여 이를 수치화한다. 가장 일반적인 방법은 이독성 공식을 적용하는 것으로 이는 글의 분량을 결정하는 단어나 문장, 문장의 구성 수준, 단어 수준 등을 통해 텍스트의 수준을 결정한다.

영어 이독성의 대표적 공식으로는 Flesch 공식(1948), Dael-Chall 공식(1958), Fry의 이독성 그래프(1977), Botel & Granowsky(1972)의 통사적 복합성 지표, Taylor(1953)의 빈칸 메우기 검사, Yngve(1960)의 문장 심도 분석 등이 있다. Flesch(1948)의 이독성 공식은 주요 기대 요인으로 4개의 요인을 선정하여 공식을 발표하였다. 100개의 단어로 된 표본에 포

함된 음절수와 한 문장 속의 단어 평균 개수로 글의 이독성을 계산하였다. 단어당 음절수는 단어의 난이도와 관련되며 문장당 단어수는 통사적 복잡성을 의미한다. Dale-Chall(1958)의 이독성 공식은 현재 가장 널리 사용되는 공식 중 하나로, 특히 3,000개 단어로 구성된 목록표를 제시하고 목록표에 등재되지 않은 단어를 중심으로 이독성을 예측한다. 대체로 이독성 공식들은 단어의 빈도, 단어의 친숙성, 단어의 길이, 문장 길이, 문장 구조의 복잡성, 텍스트의 부분적 결합 기능을 지닌 접속서나 인칭 대명사 등을 변수로 활용한다. The Gunning Fog Formula(Gunning, 1952)에서는 'foggy words'라는 3음절 이상의 단어를 주요 변수로 하여 난도를 측정하였으며 linsear Write(United Ari Force, 2014)는 문장의 길이와 3음절 이상의 단어수를 바탕으로 이독성 공식을 산출하였다(성일호, 2014: 304). 즉 이독성을 계산하기 위해 글 구성에 사용되는 기본 단위인 단어와 문장을 중심으로 단어에서 파생되는 수량적 특징이나 의미적 특징, 그리고 문장에서 파생되는 문장의 수량적 특징이나 복잡도 등을 이독성 판정에 사용한다.

공통 중핵 성취 기준(CCSS)에서 제시한 텍스트 복잡도 모형을 살펴보면 학년별 수준에 적합한 읽기 제재를 선별하는 3가지 주요 원리를 바탕으로 한다. 질적 차원, 양적 차원, 독자와 과제의 세 가지 요인으로 구성되는데 이 중 양적 차원에 해당하는 것들은 단어의 길이, 단어 빈도, 문장의 길이, 텍스트 응집성(coehsion)과 같은 글 구성 요인을 분석하여 이를 이독성 공식의 항목으로 다룬다(최숙기, 2012). 또한 서혁 외(2013)의 연구에서는 영어 이독성과 국문 이독성 공식의 선행 연구들을 분석한 결과 어휘 요인에서 단어 길이, 단어 난이도, 접속어, 인칭 대명사, 서로 다른 단어, 전치사 등을 변수로 사용하고 있으며, 문장 요인으로는 문장의 길이, 단문, 구문수치, 문장 심도 등을 변수로 사용되었다. 국문 이독성에서도 역시 어휘 요인으로 단어 난이도, 지시어, 접속어, 동음이의어, 전문 용어, 함축어, 인칭 대명사, 한자어나 외래어 등을 다루고 있고 문장 요인에서는 문장 길이,

단문, 대화문장을 변수로 선정하여 이독성 연구에 활용하고 있음을 확인할 수 있다. 지금까지의 텍스트 복잡도 또는 이독성 연구가 주로 어휘 난이도와 문장의 길이라는 두 가지 요소에 한정되어 있음을 지적하고 문장의 복잡도를 추가적으로 고려하는 방안을 제시하였다.

문장 복잡도를 수치화하여 텍스트 복잡도에 반영하는데 이때 문장 또는 통사적 복잡도란 문장이 생성되거나 변형되는 절차의 복잡성을 의미한다. 문장 복잡도의 대표적 연구로는 문장 유형의 다양성과 그 구조의 복잡성 측면을 다루고 있는 해석 문법(김의수·이로사, 2009)이 있다. 문장구조 정보를 문자와 숫자로 표시하는 선형화 모델을 구축하여, 문장의 다양성과 복잡성의 측정 기준을 8개 유형과 40가지 경우로 구분하고 각 항목에 1~25점의 복잡성 점수를 부여하는 체계를 구안하였다.

글을 읽기 어려운 정도를 판별하기 위해 글 외적 요인이나 독자 요인도 반드시 고려되어야 하지만 글 자체가 지니고 있는 내적 요인인 텍스트 요인만을 살펴보았을 때 이독성에 사용되는 요소나 방식은 작문 평가에서 사용하는 요인이나 방식과 다르지 않음을 확인할 수 있다. 국내에서 한글 텍스트에 대한 이독성 연구 사례와 각 연구에서 설정하고 있는 분석 요인을 비교하면 다음과 같다.

〈표 1〉 이독성 연구 선행연구 비교

연구자	측정 요인	측정 방법
최인숙(2005)	글자수, 단락수, 단락내문장수, 문장수, 문장내어절수, 문장내글자수, 어절수, 이형어절수, 새어절출현비율	회귀분석
이성영(2011)	글자수, 어절수, 어휘 등급의 합, 문장 수, 문장당 평균 어절 수	기술통계
김영규 외 (2011)	단어의 빈도, 추상어, 한자어, 문장 길이, 문장 구조, 대명사 개수, 접속사, 지시어, 인칭 대명사	기대 요인 선정

장미경(2011)	지시 표현, 접속 표현, 문장의 심도, 문형, 접속과 수식, 복합어, 관용 표현	자체 지수 개발
최숙기(2012)	문장수, 어려운 단어수, 펴윤 문장 길이, 평균 단어 길이	이독성 공식 적용
구민지(2013)	어휘지수, 문법지수, 문장길이, 표현지수	자체 지수 개발
서혁 외(2013)	단어 난이도, 문장 복잡도	자체 지수 개발 회귀분석
정대영(2015)	문장 복잡도	문장 복잡도 지수
조용구(2016)	쉬운 단어의 수, 평균 문장 길이	회귀분석
조찬우 외 (2018)	구문구조, 단어 난이도	자체 지수 개발
박정진(2018)	어휘 지수, 문장 길이 지수, 문법 난이도 지수, 꾸밈표현 지수	이독성 공식 적용
고승연(2018)	어휘 난도, 문법 난도, 어휘 반복, 관용구, 속담	기대 요인 선정

텍스트 이독성 또는 난이도 연구에서 주로 사용하고 있는 측정 요인은 단어와 문장의 수를 주요 요인으로 선정하고 있다. 빈도를 계산하여 난도 측정에 활용하기도 하며 이미 구축된 단어 사전을 통해 단어의 난도를 등급화하여 텍스트에 포함된 단어의 수준을 측정하기도 한다. 기본적으로 텍스트를 구성하고 있는 기본 단위인 단어의 수가 많거나 어려운 단어가 포함되어 있을 경우 어려운 텍스트로 간주하는 것을 확인할 수 있다.

그러나 단어를 기반으로 텍스트의 난이도를 측정하는 연구에서 활용하는 변인인 '단어의 빈도'나 구축된 단어 사전에서 '단어 난도'를 등급화하는 것은 현재 논의되고 있는 쉬운 공공언어 쓰기 정책에 비추어 볼 때 해당 변인을 작문 자동 채점의 변인으로 적용하기에 적절하지 않다. 자동 채점을 통해 판별하는 글의 분류는 유능한 필자의 특징을 보여야 하며 유능한 필자는 독자와의 소통을 목적으로 한다. 따라서 어려운 글에서 나타나는 특징을 좋은 글을 평가하기 위한 요인으로 활용하는 것은 글을 통한 소통의 목적에 비추어 볼 때 적절하지 않을 수 있다. 독일의 쉬운 언어 정

책(Leichte Sprache)에서 확인할 수 있는 쉬운 언어의 언어적 특징은 단순한 어휘를 사용하며, 문장을 짧게 쓰며, 동일한 대상을 표현할 때에는 동일한 단어를 사용하며, 짧은 단어를 쓸 것을 권유하고 있다(남유선, 2018). 따라서 좋은 글, 또는 잘 쓴 글의 기준을 독자와의 원활한 의사소통에 목적을 둔다면 텍스트 난도에서 활용하고 있는 변인을 잘 쓴 글의 판단 변인으로는 사용함에 있어 적절한지에 대한 고민이 요구된다.

2 문서 유사도

Abbasi & Chen(2006)은 익명성이 강한 온라인 글에서 개별 필자들의 글쓰기 양상을 파악할 수 있는 'Write-prints'(지문을 Finger-prints라고 부르는 것에서 차용)를 시각적으로 제시하기 위해 변별 요인을 구성하고 필자들의 특성을 밝히기 위한 연구를 수행하였다. 이 연구에서 나눈 변별 요인은 어휘, 통사, 구문, 내용 특성으로 큰 범주를 나누고 하위 범주로 문자 특성, 단어 기반, 구두점, 기능어 등으로 나누어 글에서 필자들의 고유 특성을 구분하여 개별 필자들의 고유한 특성을 제시하기 위한 방법을 제안하고 있다. 이처럼 글의 내용이나 특성은 필자 고유의 속성을 반영하게 된다. 이는 문체의 측면에 국한되지 않고 내용이 얼마나 유사한가에 따른 표절 판정의 영역에서 활용되기도 한다.

작문 평가에서는 모범문 또는 평가 예시문을 활용하여 글의 수준이나 등급을 나누는 기준점으로 삼는다. 이때 모범문으로 제시된 글과 얼마나 내용적 또는 형태적 유사성을 지니고 있는가에 따라 평가자는 글의 수준을 정할 수 있다. 이는 문서 유사성 측정 연구와 같은 맥락에서 이해할 수 있다. 문서 표절 또는 문서 유사도 검사는 유사성을 측정하는 방법으로 크게 형태적 유사성을 비교하는 방법과 의미적 유사성을 비교하는 방법으로 구분한다. 형태적 유사성을 비교하는 대표적인 방법으로 문장 내의 인접한 n개의 단어들을 추출하여 비교하는 n-gram 방식과 문장 내

부분 문자열(Substring)을 비교하는 방식, 벡터 공간 상에 문서를 나타내고 벡터 간의 거리 측정을 통해 유사성을 판단하는 벡터 공간 방식 등이 있다. 의미적 유사성을 비교하기 위한 방식은 단어 상호간의 의미적 관계를 계층 정보로 정의해 놓은 지식 베이스(Knowledge Base)를 이용하여 문서 내 형태적 유사성 측정 방식의 한계를 극복하기도 한다. 즉 어떤 척도에서건 텍스트 간에 유사한 특성을 공유하는 정도에 따라 동일한 범주로 분류할 수 있는지 여부를 판단할 수 있다(이은지, 2018: 11).

문자 기반 유사성 측정은 주로 문장이나 글을 이루는 단어의 형태를 비교하는 형태적 유사성 측정 방법이다. 대부분의 표절 시스템 또는 문서 유사도 비교 시스템은 기본적으로 문자 기반 유사성 측정 방식을 사용한다. 문자 기반 유사성 측정 방식은 단어의 순서를 고려하는 정확한 매칭 방법과 근사 매칭 방법으로 다시 분류된다.

정확한 매칭 방법은 두 문자열의 모든 단어가 같은 순서로 일치해야 같은 유사성 측정 대상으로 판단한다. 대표적으로 n-gram 방식이 정확한 매칭 방법이다. 이는 비교하고자 하는 두 문장 또는 문서에서 n-gram을 생성하고 추출하여, 전체에서 일치하는 n-gram의 비율로 문서 유사 여부를 판단하는 방법이다. 근사 매칭 방법은 정확한 매칭 방법의 성능 개선을 위한 방법으로 두 문장 간의 차이를 허용하여 문자열을 비교하는 방식으로, 전체 집합 속에서 두 문장 또는 두 문서 간 일치하는 부분의 비율로 유사여부를 판단하는 방식이다.

벡터 공간 모델 기반 유사성 측정은 문서를 이루고 있는 단어를 추출하여 벡터 공간 상의 벡터로 표현하고 벡터 간의 거리로 유사성을 계산한다. 문서를 이루고 있는 각 단어나 문장을 하나의 차원으로 정의하고, 만약 문서 내 단어나 문장이 포함되지 않는 경우에는 0의 값으로 표현하고, 단어나 문장이 포함된 경우는 가중치 또는 1의 값으로 표현하게 된다. 문서를 n차원의 벡터 값으로 나타낸 후, 벡터 간의 거리를 측정함으로써 문서와 문서 간의 유사성을 측정한다. 벡터 공간 모델 방식에서 유

사도 계산 방법은 다이스 유사도, 자카드 유사도, 코사인 유사도 등이 활용된다. 코사인 유사도 측정 방법이 주로 활용되는데 각각의 문서에 해당하는 열을 벡터로 놓고 두 벡터를 내적할 때, 두 벡터가 이루는 각도를 계산함으로써 유사도를 판단한다(이은지, 2018: 13).

세 번째 방법은 의미 기반 유사성 측정 방법이다. 앞선 두 가지 방법(n-gram과 벡터 공간 모델)이 형태적 유사성 측정 방법에 해당한다면 지식베이스(Knowledge base)에 정의된 개념간의 의미관계를 이용하여 지식정보를 활용한 유사성 측정이 가능하다. 이 방법의 대표적 사례로는 영어의 지식베이스인 워드넷 기반 유사성 측정 방식과 잠재 의미 분석(LSA) 방식이 있다. 워드넷은 단어의 관계를 기술하기 위해 계층적인 형태를 이루고 있으며, 상위 개념으로 올라갈수록 포괄적인 의미정보를 나타내고, 하위 개념으로 내려갈수록 구체적인 의미 정보를 나타낸다. 이러한 계층정보를 사용하여 의미적 유사성 평가가 이루어진다(이은지, 2018: 15).

〈표 2〉 문서 유사도 연구 선행 연구 비교

연구자	측정 변인	측정 방법
장성호, 강승식 (2003)	용어 선별 기법	용어 가중치 활용
장정호 외(2003)	의미 커널	헬름홀츠 머신
김혜숙 외(2003ㄱ)	단어 색인	단어 가중치
김혜숙 외(2003ㄴ)	단어와 단어쌍	빈도수
김혜숙 외(2004)	단어와 단어쌍	신경망
박선 외(2011)	단어와 유의어	유사도 알고리즘
박선영, 조환규 (2011)	성분 정렬	파라미터, 수식
박선 외(2013)	단어	의미특징 기반의 용어 가중치
강원석 외(2014)	형태소, 구문의미분석	유사도 계산 수식
강원석(2015)	용어 추출, 구문의미분석	유사도 계산 수식

문서 유사도 연구 역시 앞서 텍스트 난이도 연구와 마찬가지로 기본적인 분석 단위를 단어(형태소)로 설정하고 있다. 단어를 기반으로 가중치나 단어쌍, 빈도수 등을 활용하여 의미 유사도를 추정할 수 있는 수식으로 구성하고 각 텍스트의 의미를 수량화하여 유사도를 비교하는 방식이다. 주로 각 문서에 포함된 공통 단어의 비율을 기반으로 텍스트의 유사도를 측정한다.

텍스트 난이도 연구에서는 단어의 수나 단어 난이도를 중심으로 측정한 반면 텍스트 유사도 연구는 각 텍스트에 포함된 단어의 종류와 단어쌍의 유사도를 비교한다. 즉 텍스트에 포함된 단어들이 의미를 구성한다는 전제 하에 텍스트에 포함된 단어와 단어쌍들이 일치할수록 내용적으로 유사한 것으로 분류하는 것이다. 그러나 내용이 유사한 텍스트라 하여 좋은 글이라고 판정하는 것 역시 인간 평가자의 편향에 따른 오류를 범할 수 있다. 기존의 자동 채점 연구에서도 좋은 글의 기준을 인간 평가자의 평가 결과와 얼마나 일치하는지 여부를 통해 판단하였다. 그러나 최근의 자동 채점 연구에서는 더 이상 인간 평가자의 평가 결과를 재현하는 것을 목표로 하지 않는다(Bejar, 2011; Attali, 2013).

작문 교수·학습과 평가 영역에서 '모범문' 또는 '평가 예시문'을 통해 학습 필자들을 지도하고 평가 준거로 삼는 분야의 핵심은 동일한 내용의 글을 복제하여 쓰도록 하는 것이 아닌 기준점이 되는 글이 지니고 있는 다양한 특성들을 익히고 필자 스스로 변형하고 활용할 수 있도록 안내하는 역할을 한다(박영민, 2009). 또한 단순히 내용의 일치를 추구하는 것이 아니라 조직, 구성, 표현 등 다양한 글의 특성을 학습 필자들이 익히도록 하는 것에 목적이 있기 때문에 단어나 단어쌍의 일치 여부를 통해 좋은 글을 판단하는 기준으로 삼는 것 역시 한계를 지닌다.

3. 자동 작문 채점 연구

Page(1966)의 에세이 채점 자동화 시스템을 시작으로 자동 채점에 대한 많은 연구가 인간 평가자의 평가 결과와 자동채점 결과의 상관관계를 통해 가능성을 검증해왔다. 주로 텍스트에서 변별적, 분절적으로 확인할 수 있는 텍스트 요인을 추출하여 알고리즘을 구축하는 선형 모델링 접근법을 사용한다(Attali & Burstein, 2006; McNamara, Crossley, & Roscoe 2013). 최근에는 인가 평가자에 의해 평가된 글에 대해 기계 학습한 결과를 바탕으로 추가적인 텍스트에 대한 자동채점이 이루어진다.

Page(1966)가 개발한 PEG(Project Essay Grade)는 1998년 웹상에서 평가가 가능하도록 구축되었다. 이 시스템은 'trins'와 'proxes'를 측정하는데 trins은 구두점, 유창성, 문법 등과 같은 상위 변수로 이는 직접 측정될 수 없으며 다른 측정 변수에 의해 근사치가 계산된다. 이 trins를 계산하기 위한 실제 변수들을 proxes라고 한다. 예를 들어 유창성을 개념 준거로 산정하여 개념 준거를 측정하기 위한 실제 준거를 마련하는 것이다. E-rater는 Educational Testing Service(Burnstein, 1998)에서 개발한 자동 채점 시스템이다. 문법 오류, 단어 사용 오류, 철자법 오류, 담화 요소, 문체상의 특징, 주제별 단어 사용을 측정하기 위한 벡터 분석, 단어의 상대적 빈도, 단어의 정교함, 단어의 배열 등을 고려한다. 회귀모델을 사용하여 자동 채점 결과를 산출한다. IEA는 Pearson Knowledge Technologies(Foltz, Laham, & Landauer, 1999)에서 개발하였다. 이 시스템은 잠재 의미 분석과 기계 학습을 결합한 방식을 사용한다. 내용(잠재 의미 분석, 벡터 공간 분석), 단어(단어 사용의 유형, 다양성, 다의어의 사용 등), 문법(n-gram, 문법 오류의 수와 유형), 맞춤법, 조직, 문체 등을 다각도록 분석하여 자동 채점을 수행한다. 앞선 연구들과 마찬가지로 인간 평가자의 평가결과를 학습한 후 추가 텍스트에 대한 평가가 이루어진

다. 이 외에도 다양한 자동채점 시스템이 구축되어 있으며 텍스트에서 추출할 수 있는 여러 가지 변수들의 유형과 조합 방식을 통해 자동채점이 이루어진다.

〈표 3〉 자동 채점 연구 선행 연구 비교

자동 채점 시스템	측정 변인	측정 방법
Lexile Writing Analyzer(Smith, 2009)	의미의 복잡성 통사적 정교함	렉사일 척도 인간 평가 결과 필요하지 않음
LightSIDE (Mayfield & Rose, 2010)	n-gram POS 태깅과 수량화	선형 회귀 Naïve Bayes 선형 서포트 벡터 머신
CRASE (Lottridge et al., 2013)	아이디어, 문장 유창성 조직, 태, 단어 선택, 담화 관습 등	기계 학습
Intellimetric (Schultz, 2013)	내용 속성(단어의 다양성, 응집력, 일관성 등) 구조 속성(문법, 맞춤법, 문장 완성도, 문장 다양성, 복잡성, 주어-동사 일치 등)	선형 분석 베이지안 접근법 잠재 의미 분석 등

그러나 이러한 자동채점에 대한 연구에서 몇 가지 문제점으로 인해 새로운 방향으로의 자동채점 연구가 진행되고 있다. 먼저 인간 평가자의 결과에 대한 기계학습 결과를 적용하는 방식에 있어서 편향이 발생할 수 있다. 체계적 혹은 비체계적인 인간 평가자의 오류로 인해 발생하는 잠재적 오류를 극복하기 위해 사전에 평가된 결과 없이 텍스트에서 나타나는 오류를 탐지하거나 인간 평가자의 판단과 별개로 자연언어처리나 기계 학습을 통해 확인할 수 있는 좋은 글의 특징을 탐색하여 채점이 이루어진다. 또한 텍스트의 언어적 특성에 주로 초점을 맞추어왔던 방식에서 벗어나 텍스트의 의미를 자동채점의 변인으로 활용하기 위한 시도들이 이루어지고 있다. Chali & Hasan(2013)은 자동 채점 분야에서 활용하는

잠재 의미 분석이 지닌 한계를 지적하였다. 잠재 의미 분석은 단어의 순서나 통사적 의미 구조를 확인할 수 없고 단지 단어의 빈도만을 측정할 수 있다. 따라서 이를 극복하기 위해 Semantic Tree Kernel을 자동 채점에 적용하여 텍스트의 각 문장을 분석하여 구문 유사성을 계산하였다. 또한 Taghipour & Ng(2016)와 Alikaniotis et al.,(2016)은 신경망을 자동 채점에 적용하였다. 텍스트의 특정 요인을 추출하여 선형적 접근을 하는 것이 아니라 입력된 텍스트와 배정된 점수 간의 관계를 스스로 학습하여 자동 채점에 필요한 정보를 인코딩하여 텍스트가 지닌 데이터들의 복잡한 패턴을 학습하게 하는 방식이다.

작문 자동 채점은 채점을 위한 텍스트의 형식적 특성과 내용적 특성을 모두 고려할 수 있어야 하며 그 특성들을 결합하는 방식에 따라 자동 채점 방식을 다양화할 수 있다. 특히 최근 논의되고 있는 신경망을 활용한 자동 채점은 선형적 접근으로 이루어져 온 자동 채점의 방식을 확장시켰다는 점에서 의의가 있다. 선형적 접근은 좋은 평가를 받는 글이 모두 유사한 또는 단일한 특징을 지니고 있다고 가정한다. 그러나 좋은 평가를 받는 글은 각기 다른 변별적 특성에 의해 설명될 수 있다. 이처럼 하나의 기준점을 적용하여 글에 대한 평가를 내리는 방식이 아니라 텍스트의 특성들이 변별적으로 결합된 다양한 기준점을 적용하여 글의 수준을 분류할 수 있는 클러스터 방식(Crossely, Roscoe, & McNamara, 2014)도 활용된다.

Ⅳ. 제언

작문 평가에서 상황이나 평가자로 인해 발생하는 어려운 점들을 극복하고 보조 또는 대안의 수단에서 자동 작문 평가의 연구가 활발히 진행

되고 있다. 기본적으로 인간 평가자는 수량화할 수 없는 글을 읽고 이미지화하여 비교, 판단하는 과정을 거친다면 기계식 평가의 경우 그 동안 수량화하지 못했던 글의 여러 요인을 척도화하여 이를 형태 또는 의미 분석에 사용하는 단서로 삼는다. 따라서 작문평가에서 평가의 정확성과 일관성의 문제, 그리고 평가 결과의 객관적 제시의 문제에 있어 자동 평가는 시간과 비용의 절약이라는 실제적인 문제에서부터 평가 과정과 결과의 측면에서도 작문 평가에 많은 개선을 가져올 것이다.

작문 평가를 위한 평가 준거와 기준의 마련에서 준거에 대한 정확한 기준 마련과 기준 해석의 어려움이 있어 왔으며 자동 채점을 위한 수많은 연구에서 텍스트의 어떠한 척도들이 정확한 작문 평가를 위한 척도로서의 역할을 할 수 있을지에 대한 고민을 쏟아내고 있다. 자동 채점의 고민은 결국 '잘 쓴 글을 어떻게 측정하고 수량화할 수 있는가?'에 대한 답을 제시하는 것이다. 텍스트가 지니고 있는 여러 형태적, 의미적 정보 중에서 무엇을 척도로 삼아 수리적 변형을 통한 결과 예측의 정확성과 설명력을 높일 수 있는 것인지가 관건이 된다. '잘 쓴 글' 또는 '유능한 필자의 특성'이라는 추상적인 목표에 대해 어떠한 개념 준거와 실제 준거를 적용할 것인지에 대한 고민이 요구된다. 개념 준거는 기존의 인간 평가자들이 활용한 준거와 크게 달라지지 않을 것이다. 이는 오랜 기간 이루어져 온 작문에 대한 연구의 결과물이며 교육과정에서 학생들이 성취해야 하는 기준으로 다루어지고 있기 때문이다. 그렇다면 개념 준거를 정확히 설명하고 포괄할 수 있는 실제 준거에 대한 고민이 필요하다. 하나의 척도를 선정하여 글을 분석하고 결과를 도출하였을 때 인간 평가자들의 결과와 상관관계가 높다는 것에서 그칠 것이 아니라 해당 척도가 온전한 실제 준거의 역할을 다할 수 있을지에 대한 고민을 통해 단순히 측정이 아닌 평가의 범주로 나아갈 수 있는 기틀을 마련해야 한다.

현재 이루어지고 있는 자동 채점 연구의 방식이 텍스트에 포함된 텍스트 요인들의 수량적 정보에 초점을 두고 있는 것은 독자의 이해와 텍

스트의 활용도를 높이기 위한 쉬운 언어 정책과도 배치되는 부분이 있다. 이는 자연언어처리를 통한 텍스트 분석의 수많은 분야 중에서 국어교육 연구 영역에서도 다루어져 왔던 이독성 연구, 문서 유사도 연구들과 비교하고 고민해야할 지점이 발생한다.

텍스트 분석 연구는 모두 같은 방법과 절차를 공유하고 있다. 이독성이나 문서 유사도, 그리고 작문 평가 연구 모두 수량화할 수 있는 척도를 통한 글의 의미, 수준, 유사도 등을 계산한다. 같은 척도를 사용한다 하더라도 해당 척도를 어떻게 바라볼 것인가에 대한 고민이 요구된다. 이독성 연구, 문서 유사도 연구와 비교하여 작문 평가 연구가 분화하고 변별되어야 하는 지점은 어디인지에 대한 논의와 연구를 통해 작문 평가가 인간 평가를 보조 또는 대체함에 있어서 작문 평가에서 그동안 미흡했던 또는 불가능했던 영역에 있어 어떠한 정보를 제공할 수 있으며 작문 평가 영역에서 어떠한 가치를 지닐 수 있을 것인가에 대한 고민이 선행되어야 할 것이다.*

* 이 논문은 2019년 05월 20일에 투고되었으며 6월 19일에 심사 완료되어 6월 21일에 게재 확정되었음

참고문헌

강원석·황도삼(2014), 「구문의미분석를 이용한 유사문서 판별기, 『한국콘텐츠학회 논문지』 14(3), 40-51, 한국콘텐츠학회.

강원석(2015), 「구문의미트리 비교기를 이용한 유사문서 판별기」, 『한국콘텐츠학회논문지』 15(10), 636-646, 한국콘텐츠학회.

고승연(2018), 「이독성 공식 개발을 위한 어휘 목록과 문법 항목의 기대 요인 선정」, 『한국어문화교육』 11(2), 1-24, 한국어문화교육학회.

구민지(2013), 「한국어 읽기 교육을 위한 텍스트 난이도 측정법 연구」, 카톨릭대학교 박사학위논문.

구경호(2015), 「평가기준의 설정 및 활용」, URL: http://contents.kocw.or.kr/KOCW/document/2015/pusan/kookyungho/8.pdf

김영규·이승은·이지은(2009), 「상세화를 통한 한국어 텍스트의 이독성 향상 방안 연구」, 『이중언어학』 41, 57-81, 이중언어학회.

김의수·이로사(2009), 「교육과정에 따른 문장의 다양성과 복잡성 추이」, 『한국언어문학』 69, 83-115, 한국언어문학회.

김혜숙·박상철·김수형(2003ㄱ), 「단어가중치 기반 문서간 유사도 측정에 관한 연구」, 『한국멀티미디어학회 학술발표논문집』 198-201, 한국멀티미디어학회.

김혜숙·박상철·김수형(2003ㄴ), 「단어 및 단어쌍 별 빈도수를 이용한 문서간 유사도 측정」, 『대한전자공학회 학술대회자료집』, 1311-1314, 대한전자공학회.

김혜숙·박상철·김수형(2004), 「단어/단어쌍 특징과 신경망을 이용한 두 문서간 유사도 측정」, 『정보과학회논문지: 소프트웨어 및 응용』 31(12), 1660-1671, 한국정보과학회.

남유선(2018), 「독일의 '쉬운 언어 Leichte Sprache'에 나타나는 언어적 특징과 활용 사례」, 『독일언어문학』 82, 101-114, 한국독일언어문학회.

박선·김경준·이진석·이성로(2011), 「군집 주제의 유의어와 유사도를 이용한 문서 군집 향상 방법」 『전자공학회논문지-SP』 48(5), 30-38, 대한전자공학회.

박선·김경준·김경호·이성로(2013), 「의미특징 기반의 용어 가중치 재산정을 이용한 문서군집의 성능 향상」, 『한국정보통신학회논문지』 17(2),

347-354, 한국정보통신학회.
박선영·조환규(2011), 「성분 정렬을 이용한 한글 유사 문서 탐색 방법」, 『한국정보과학회 학술발표논문집』 38(1C), 228-231, 한국정보과학회.
박영민(2009), 「평가 예시문을 활용한 쓰기 평가 개선 방안」, 『청람어문교육』 39, 111-133, 청람어문교육학회.
박영민·이재기·이수진·박종임·박찬흥(2016), 『작문교육론』, 역락.
박정진(2018), 「이독성을 활용한 한국어 읽기 자료의 수준 설정 가능성 탐색」, 『현대사회와다문화』 8(2), 1-22, 대구대학교 다문화사회정책연구소.
서혁·이소라·류수경·오은하·윤희성·변경가·편지윤(2013), 「읽기(독서) 교육 체계화를 위한 텍스트 복잡도 상세화 연구 (2)」, 『국어교육학연구』 47, 253-290, 국어교육학회.
성일호(2014), 「이독성공식과 Coh-Metrix 를 활용한 우리나라 고등학교 영어교과서 이독성 분석」, 『영어영문학연구』 40(4), 299-320, 한국중앙영어영문학회.
오세영(2014), 「쓰기 평가자의 정보 처리 과정 연구-채점 과정의 정보 처리 모델 비교를 중심으로」, 『한어문교육』 31, 109-154, 한국언어문학교육학회.
이성영(2011), 「읽기 (독서) 에서의 교육 내용 위계화: 초등 교과서의 이독성 비교 연구-국어, 사회, 과학 교과서를 중심으로」, 『국어교육학연구』 41, 169-193, 국어교육학회.
이은지(2018), 「확장된 의미역 결정을 이용한 문서 유사성 판단」, 조선대학교 박사학위논문.
장미경(2011), 「한국어 읽기 교육을 위한 텍스트 난이도 평가 방안 연구」, 고려대학교 박사학위논문.
장성연(2017), 『디자인 평가의 제한적 한계의 문제와 확장적 평가를 위한 고찰』, 『브랜드디자인학연구』 15(3), 127-138, 한국브랜드디자인학회.
장성호·강승식(2003), 「용어 선별 기법에 의한 유사 문서 판별 시스템」, 『한국정보과학회 학술발표논문집』 30(1B), 534-536, 한국정보과학회.
장정호·김유섭·장병탁(2003), 「헬름홀츠머신 학습 기반의 의미 커널을 이용한 문서 유사도 측정」, 『한국정보과학회 학술발표 논문집』 30(1), 한국정보과학회.
정대영(2015), 「소설 텍스트의 문장 복잡도 연구-자동화된 프로그램을 활용하여」, 『문학교육학』 48, 263-292, 한국문학교육학회.
조용구(2016), 「글의 수준을 평가하는 국어 이독성 공식」, 『독서연구』 41, 73-92, 한국독서학회.
조찬우·조찬형·우균(2018), 「가독성 평가를 위한 한글 문장의 복잡도 측정 방

법」, 『한국정보과학회 학술발표논문집』 2265-2267, 한국정보과학회.
최숙기(2012), 「텍스트 복잡도 기반의 읽기 교육용 제재의 정합성 평가 모형 개발 연구」, 『국어교육』 139, 451-490, 한국어교육학회.
최인숙(2005), 「독서교육시스템을 위한 텍스트수준 측정 공식 구성에 관한 연구」, 『정보관리학회지』 22(3), 213-232, 한국정보관리학회.
Abbasi, A., & Chen, H.(2006, May), Visualizing authorship for identification. *In International Conference on Intelligence and Security Informatics* (pp. 60-71). Springer, Berlin, Heidelberg.
Alikaniotis, D., Yannakoudakis, H., & Rei, M.(2016), *Automatic text scoring using neural networks*. arXiv preprint arXiv:1606.04289.
Attali, Y., & Burstein, J.(2006), Automated essay scoring with e-rater® V. 2. The Journal of Technology, *Learning and Assessment*, 4(3).
Attali, Y.(2013), 11 Validity and Reliability of Automated Essay Scoring. *Handbook of automated essay evaluation: Current applications and new directions*, 181.
Baird, J. A.(2000), Are examination standards all in the head? Experiments with examiners' judgements of standards in A level examinations. *Research in Education*, 64(1), 91-100.
Bejar, I. I.(2011), A validity-based approach to quality control and assurance of automated scoring. Assessment in Education: Principles, *Policy & Practice*, 18(3), 319-341.
Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M.(1998, April), Computer analysis of essays. In *NCME Symposium on Automated Scoring*.
Chali, Y., & Hasan, S. A.(2013), On the effectiveness of using syntactic and shallow semantic tree kernels for automatic assessment of essays. *In Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 767-773).
Crisp, V.(2012), An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice,* 31(3), 10-20.
Crossley, S. A., Roscoe, R., & McNamara, D. S.,(2014), What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication*, 31(2), 184-214.
Eckes, T.(2012), Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, 9(3), 270-292.
Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M.(2017), Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*.

Foltz, P. W., Laham, D., & Landauer, T. K.(1999), The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic *Journal of Computer-Enhanced Learning*, 1(2), 939-944.

Freedman, S. W., & Calfee, R. C.(1983), Holistic assessment of writing: Experimental design and cognitive theory. *Research on writing: Principles and methods*, 75-98.

Golding, C., Sharmini, S., & Lazarovitch, A.(2014), What examiners do: What thesis students should know. *Assessment & Evaluation in Higher Education*, 39(5), 563-576.

Griffin, T. D., Jee, B. D., & Wiley, J.(2009), The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition*, 37(7), 1001-1013.

Kawabata, H., & Zeki, S.(2004), Neural correlates of beauty. *Journal of neurophysiology*, 91(4), 1699-1705.

Leder, H., Belke, B., Oeberst, A., & Augustin, D.(2004), A model of aesthetic appreciation and aesthetic judgments. *British journal of psychology*, 95(4), 489-508.

Lottridge, S. M., Schulz, E. M., & Mitzel, H. C.(2013), Using automated scoring to monitor reader performance and detect reader drift in essay scoring. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, MD Shermis and J. Burstein, Eds. New York: Routledge, 233-250.

Mayfield, E., & Rosé, C. P.(2010), An interactive tool for supporting error analysis for text mining. *Proceedings of the NAACL HLT 2010 Demonstration Session*, 25-28.

McNamara, D. S., Crossley, S. A., & Roscoe, R.(2013), Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods*, 45(2), 499-515.

Muchinsky, P. M., Culbertson. S. S.(2016), 『산업 및 조직 심리학』, 유태용 역, 시그마프레스.

Page, E. B.(1966), The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5), 238-243.

Pressley, M., Borkwski, J. G., & Schneider, W.(1989), Good information processing: What it is and how education can promote it. *International Journal of Educational Research*, 13(8), 857-867.

Schultz, M. T.(2013), The IntelliMetric automated essay scoring engine-a review and an application to chinese essay scoring. *Handbook of automated essay*

scoring: Current applications and future directions, 89-98.

Smith, C., & Dunstan, A.(1998), Grade the learning, not the writing. *The theory and practice of grading writing: Problems and possibilities*, 163-70.

Smith, M.(2009), The Reading-Writing Connection. Technical report, MetaMetrics, 2009. URL https: //metametricsinc.com/research-publications/reading-writing-connection/.

Taghipour, K., & Ng, H. T.(2016), A neural approach to automated essay scoring. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882-1891).

투 고 자 : 이 현 준
근 무 지 : 중원대학교
직 위 : 초빙교원
연 락 처 : 043-830-8393
전자우편 : 9th2out_flow@naver.com

투 고 자 : 박 영 민
근 무 지 : 한국교원대학교
직 위 : 교수
연 락 처 : 043-230-3547
전자우편 : enrapture@knue.ac.kr

■ 국문초록 ■

자연언어처리를 활용한 텍스트 연구 비교를 통한 자동채점 변인 탐색 연구

이현준 · 박영민
중원대학교 초빙교원 · 한국교원대학교 교수

이 연구는 비정형 데이터인 텍스트에 대해 컴퓨터를 활용하여 글의 수준을 판별하기 위해 활용할 수 있는 변인 탐색을 목적으로 하였다. 이를 위해 텍스트 분석과 유사한 분야인 미술 작품에 대한 신경미학을 검토하고 미술 분야에서 활용한 인공지능 알고리즘을 검토하였다.

또한 자동 채점에서 활용할 수 있는 텍스트 요인과 분석 방법을 탐구하고자 자연언어처리를 활용한 텍스트 분석 연구의 여러 분야를 비교하였다. 각 연구 분야별로 활용하고 있는 텍스트 분석 요인과 분석 방법을 살펴봄으로써 자동 채점과의 유사점과 차이점을 도출하고자 하였다.

무엇보다 잘 쓴 글, 또는 좋은 글에 대한 명확하고 합의된 정의가 요구된다. 이를 통해 글의 수준을 설명할 수 있는 수량화, 정량화 가능한 요인을 추출할 수 있다. 또한 텍스트를 구성하고 있는 기본 요소인 단어(형태소)가 텍스트 분석의 기본 요인이지만 단어가 가진 어떤 속성을 활용할 것인지는 연구 분야마다 차이를 보인다.

텍스트는 단어의 집합으로 구성되지만 텍스트의 의미는 단어의 단순 총합 이상이다. 따라서 텍스트의 특성을 설명할 수 있는 단어 이외의 변인을 선정할 수 있어야 하며 각 변인은 텍스트의 특성을 정확하고 변별적으로 측정할 수 있어야 한다.

〈**핵심어**〉 : 자동 채점, 텍스트 분석 요인, 이독성, 문서 유사도

■ ABSTRACT ■

A Study on the Search for Automatic Scoring Variables by Comparison of Text Studies Using Natural Language Processing

Lee Hyeon-Jun · Park Young-Min

Lecturer, JungWon University · Professor, Korean National University of Education

The purpose of this study was to explore variations that could be used to determine the level of text by using computers which is unstructured data. For this purpose, we examined the neural aesthetics of art works similar to text analysis and examined artificial intelligence algorithms used in the field of art.

In addition, we compared several areas of text analysis research using natural language processing to explore textual factors and analysis methods that can be used in automatic scoring. By looking at the text analysis factors and analysis methods used for each research area, we tried to draw similarities and differences from automatic scoring.

Above all, a clear and agreed definition of good writing is required. In this way, quantifiable factors that can explain the level of text can be extracted. In addition, words (forms), the basic elements that make up the text, are the primary factors in text analysis, but whether a word will take advantage of any attribute it has varies from field to field of study.

Text consists of a set of words, but the meaning of the text is more than the simple sum of words. Therefore, we should be able to select variables other than words that can describe the nature of the text. Each variable should be able to accurately and selectively measure the characteristics of the text.

〈Key words〉 : Automatic grading, text analysis factors, text difficulty, document similarity