쓰기 평가의 키워드 채점 방식에 대한 타당성 분석

이지원 한국교원대 국어교육 박사과정

I. 서론

쓰기 평가는 평가 전반에 걸쳐 학습자의 수행을 요구하며, 그러한 수행의 결과물인 '글'을 놓고 평가가 이루어진다. 글의 수준에 대한 평가 결과를 도출하기 위해서는 평가자의 판단에 의한 채점 과정이 필요한데, 이는 인간 평가자의 심적 판단에 따른 평가의 본질상 오래전부터 객관성의 문제에서 자유로울 수 없었다. 같은 글에 대한 평가에서 평가자 간에 점수가 다양하게 나타나거나(Dunbar *et al.*, 1991), 심지어 한 평가자가 같은 글을 재채점하는 경우에도 다른 점수가 부여되는(Eells, 1930) 경향이 나타나기 때문이다.

쓰기 평가에서 평가자는 측정 도구이자, 점수 부여에 대한 의사결정을 내리는 평정자로서 높은 전문성이 요구되는 평가 요인이 된다. 쓰기 평가의 평가자 요인이 평가 결과에 미치는 영향에 대하여 평가자의 후광 효과(Thorndike, 1920)를 비롯하여, 채점 척도 사용의 중앙집중경향(Landy & Farr, 1983), 평가자의 엄격성과 일관성(Saal *et al.*, 1980), 채점 순서(Hughes & Keeling, 1984) 등이 평가 결과에 영향을 미치는 평가자 요인으로 지적되어 왔다.

그런데 쓰기 평가에서 수행 결과에 대한 평가자 간의 오차가 발생하는 경우, 평가의 신뢰도뿐만 아니라 타당도 또한 보장할 수 없다. 또한 최근 들어 대단위 고부담 평가에서 쓰기 평가의 영역이 점차 확대되고 있는 추세이므로, 공정하고 객관적인 평가에 대한 사회적 기대에 부응하기 위해서는 이러한 평가자 간 점수의 오차를 줄일 수 있는 채점 방안이 마련되어야 한다.

1. 평가자 오차 요인의 최소화

평가자 요인이 평가 결과에 미치는 오차를 최소화하기 위하여 평가자 훈련, 평가자 간 협의, 평가 기준의 상세화, 평가 예시문의 사용 등의 노력을 기울인다 하더라도 평가자의 주관성을 완전히 배제하기란 사실상 어렵다. 따라서 최근에는 쓰기 평가에서 다국면 Rasch 모형과 같은 통계적 방법에 의하여 원점수를 조정하거나 부적격한 평가자의 점수를 배제하는 방법이 제안되었으며(Linacre, 1990), 더 나아가 한국어 서답형 문항의 컴퓨터 자동 채점 방안(노은희 외, 2012)이 연구되어 오기도 했다.

쓰기 평가에 드는 시간과 노력에 따른 비용 및 효율성 또한 대단위 평가에서 쓰기 평가가 저변을 확대하지 못하는 원인 중 하나이다. 여러 편의 글을 채점하는 작업은 적지 않은 시간과 노력을 요구하며, 평가자의 피로 누적 및 이에 따른 평가자 내 신뢰도(intra-rater reliability)의 저하가 일어나기 매우 쉽다.

쓰기 평가에서 이러한 평가자 신뢰도 저하와 비용 손실을 최소화하고 객관성을 기할 수 있는 채점 방식의 하나로 키워드 채점이 있다. 키워드(keyword)는 문서를 대표하는 단어들의 집합으로 정의될 수 있는데 이 키워드는 문서의 내용을 적절히 요약, 반영한다(김일환 · 이도길, 2011). 키워드 채점(keyword matching scoring)이란 주어진 문항에서 요구되는 키워드를 사전에 선정하고, 이에 따라 답안에서 키워드가 일치되었을 때 정답으로 판정하여 점수에 산정하는 채점 방식이다.

인간 평가자의 주관적 판단을 배제하면서 글의 질을 판단할 수 있는 요인은 매우 다양한데, 지금까지는 문장의 길이, 단어 빈도와 같이 수량적인 방식으로 문장의 성숙도(syntactic maturity)를 판단하는 언어 기반 평가(Language-based Assessment)가 대표적이었다. 글에 대한 양적 분석이 가능하다는 점은, 이러한 결과의 통계적 처리를 통해 데이터를 얻고 데이터의 처리를 통한 여러 가지 분석이 가능하다는 이점으로 자연스럽게 연결된다. 그러나 단순히 문장의 성숙도만을 기준으로 채점이 이루어지는 경우, 쓰기 과제의 특성을 반영할 수 없고 글의 표층적 수준에 대한 측정만 가능하다는 단점이 있다. 반면 특정 쓰기 과제에서 요구되는 키워드를 추출하고, 이에 따라 채점을 하게 되면 채점 결과에 대한 양적 분석이 가능할 뿐만 아니라, 출제자가 의도한 쓰기 과제의 평가 목적에 따라 글에 적절한 키워드가 포함되었는지에 대해 파악할 수 있다는 장점이 있다.

또한 키워드 채점 방식은 쓰기 평가에서 가장 큰 난제 중 하나인 평가자의 주관성이 배제된다는 장점이 있다. 키워드에 의한 평가 기준이 세워지고 나서는 객관적이고 기계적인 방법으로 채점이 수행되기 때문에 평가자 간 이견이 발생할 가능성이 거의 없기 때문이다. 또한 평가자가 글을 읽는 데 들이는 인지적 부담과 비용을 줄이고 신속한 채점을 가능케 하기 때문에 특히 대규모로 쓰기 평가가 이루어지는 경우, 현재로서는 대부분의 시험에서 키워드 채점 방식이 적용되고 있다.

2. 키워드를 통한 집단 간 특성 변별

코퍼스 언어학에서는 각 코퍼스에서 추출된 키워드를 통해 코퍼스 간의 차이점을 설명한다. 영국 국립 코퍼스(British National Corpus)를 대상으로 한 연구에서 Paquot & Betgen(2008)은 하위 코퍼스인 'academic'과 'fiction' 범주를 비교 분석하여 나타나는 키워드의 차이를 보여 주었으며, Leech *et al.*(2001)은 말하기와 쓰기 상황에서 나타나는 코퍼스 간 차이를

검증하면서 하위 코퍼스인 'imaginative writing'와 'informative writing'를 비교하여 단어 빈도 간에 차이가 나타남을 밝혔다. 이는 발화 영역 및 담화 상황에 따라 서로 다른 단어 사용 양상이 나타남을 의미한다.

특히 같은 언어를 사용하는 영어 사용자들 간에도 문화권에 따른 단어 사용 빈도의 차이가 나타남을 보여 준 Oakes & Farrow(2007)의 연구 및 각 학문 분야에 따른 차이를 분석한 Hyland & Tse(2007)와 같이 키워드는 각 언어 사용자 집단에 대한 특성을 드러내는 지표라 할 수 있다. 보다 구체적으로 Hyland & Tse(2007: 238)는 3,300만 단어의 학술어 목록(Academic Word List, ASW, Coxhead, 2000) 코퍼스를 대상으로 한 분석에서 각 학문에 대한 하위 코퍼스(엔지니어링 · 과학 · 사회과학)의 570개의 집단 중 534(94%)개 집단이 분산의 이질성을 보였다. 이는 "모든 학문들은 그 학문에서 사용되는 단어들을 형성(2007: 240)"하며, 같은 단어라 할지라도 빈도 및 사용 양상이 다르다는 점을 시사한다. 또한 하위 코퍼스 간 연어 분석을 통해서 단어 'strategy'가 경영학 분야에서는 'marketing strategy', 응용 언어학 분야에서는 'learning strategy', 사회학 분야에서는 'coping strategy'와 같이 학문 간 연어 발생 양상 및 빈도에서 차이가 나타남을 보여 준 바 있다.

즉 사용역(register)에 따라 수집된 코퍼스에서 추출된 키워드는 각 집단의 특성을 명확하게 드러낸다고 볼 수 있다. "어떤 단어가 한 코퍼스의 특징을 드러내는가(Kilgarriff, 2001)"를 보여 주는 정도를 '키워드성(keywordness)'(Witten *et al.*, 1999)이라 하는데 이 키워드성이 높을수록 해당 집단의 특성을 나타내는 단어로 볼 수 있다. 이를 쓰기 평가 국면으로 가져오면, 쓰기 수행에 따라 집단을 수준별로 구분하고, 이에 따라 각 집단을 대표하는 키워드를 추출하여 같은 쓰기 과제에 대해 각 집단이 어떠한 양상으로 반응하였는지를 탐색할 수 있다. 이를 통해 키워드가 각 집단의 쓰기 수행에서 드러나는 특성을 파악할 수 있다. 특정 글(답안)이 어느 수준에 도달했는지 판가름하는 기준으로써 키워드가 대표성을 지닌다고 판단하는 것이다.

3. 쓰기 평가의 컴퓨터 채점

키워드 채점은 더 나아가 쓰기 평가의 컴퓨터의 자동 채점을 가능케 하는 토대가 된다. 이미 영미권에서는 1960년대부터 자동 채점 프로그램의 개발이 이루어져 왔으며, ETS 주관의 쓰기 평가에서는 e-rater(Attali & Burstein, 2006)를 활용한 컴퓨터 자동 채점 방식이 보편화되어 있다. 현재 우리나라에서도 한국어로 작성된 답안의 자동 채점을 위한 프로그램이 개발 단계에 놓여 있다.

컴퓨터 자동 채점 시스템인 ETS(Education Testing Service)의 e-rater와 LSA(Latent Semantic Analysis)를 활용한 IEA(Intelligent Essay Assessor)(Landauer, Laham & Foltz, 2003)은 모두 글 단위의 데이터를 처리하는 에세이 평가 프로그램인데, 분석 단위의 기본은 단어이다. 쓰기 평가의 자동 채점 프로그램은 자연어 처리(Natural Language Processing) 기술을 바탕으로 하는데, e-rater는 이러한 자연어 처리 기술을 바탕으로 크게 통사적 특질 · 담화구조 · 화제 분석의 세 가지 요소로 평가 결과를 도출한다.

첫째, 통사적 특질의 경우, 글의 동사 및 절의 유형이 무엇인지에 따라 채점되는데, 이는 동사와 절 표지어를 통해 통사적 다양성을 판단함으로써 평가된다.

둘째, 담화 구조의 경우 논항 분석에 사용되는 큐워드(cue word)에 의존하여 채점된다. 예를 들면 논항 전개에서 'perhaps', 'possibly'는 필자의 신념을 나타내는 단어(belief word)로 간주되며, 'this', 'these'는 동일한 화제를 전개할 때 사용하는 표지이다.

셋째, 화제 분석에서는 화제에 대한 논의의 전개에서 수험생의 수준을 단어에 근거를 두고 판단한다. 좋은 글은 부족한 글보다 화제에 대한 논의가 더 구체화되고 정확한 단어가 사용되는 경향이 있다. 따라서 좋은 글이 다른 좋은 글과 단어 선택이 유사한 면이 있으며, 반면 부족한 글은 다른 부족한 글의 단어 선택과 유사하다고 기대한다. 이에 따라 e-rater는 에세이의 화제

내용을 각 6개 채점 요소를 통해 인간평가자에 의해 수동적으로 채점된 평가자 훈련용 표집 글에서 추출된 단어들을 비교함으로써 평가되는데, 이는 단어 빈도 및 빈도에 의한 가중치에 근거하여 계산된다(Burstein, 2001). 즉, e-rater의 화제 분석 채점은 글을 잘 쓴 수험생들이 많이 쓴 단어의 빈도를 중심으로 평가되도록 채점 프로그램이 설계되어 있다.

키워드 채점은 위와 같이 현재 시행되고 있는 컴퓨터 채점 방식의 원리와 유사하게 단어를 기반으로 둔 채점 방식이다. 기존의 인간 평가자에 의한 채점보다 시간과 비용이 절감되고 보다 객관적이지만, 이에 따른 채점의 결과가 쓰기 평가에서 일반적인 채점 결과에 따른 점수와 같이 수험생의 능력을 얼마나 측정하는지는 밝혀진 바가 없다. 따라서 이 연구에서는 인간 평가자의 채점 결과에 따라 분류된 수준별 집단에서 키워드를 자동적으로 산출하여 분석함으로써, 키워드가 각 집단의 수준을 드러내는지의 여부를 탐색하고자 한다.

최근 국어과에서는 고등사고능력의 교수 · 학습과 이에 따른 평가의 중요성이 대두됨에 따라, 대단위 쓰기 평가의 도입에 대한 필요성이 점차 증대되고 있다. 사실 대단위 쓰기 평가에 대한 높은 교육적 요구가 있어 왔지만, 이것이 현실적으로 시행되지 못했던 원인에는 시행 과정에서의 채점에 소요되는 시간적 · 경제적 · 행정적 부담이 가장 큰 원인으로 지적되고 있다(노은희 외, 2012: 7). 따라서 대단위 쓰기 평가가 보편화되기 위해서는 채점 방식의 간소화와 자동화가 선행적으로 이루어져야 한다. 이 시점에서 키워드 채점 방식은 컴퓨터 자동 채점과 인간 평가자의 채점 사이에서 대안적 채점 방식으로써 활용이 기대되고 있다.

만약 키워드가 인간 평가자에 의해 신뢰롭게 채점된 결과에 따라 구분된 각 집단에 따라 서로 다른 양상 및 특성을 보인다면, 학생의 쓰기 수준을 판단하는 것에 대한 타당성을 확보하였다고 말할 수 있다. 이를 통해 키워드 채점 방식이 쓰기 평가 채점 방식의 하나로 자리매김할 수 있을 것이다.

4. 연구 질문

이 연구는 텍스트 기반의 키워드 자동 산출 방법을 통해 키워드가 쓰기 평가에서 수험생의 수준을 얼마나 타당하게 판별할 수 있는지를 탐색하기 위하여 수행되었다. 이를 위해 인간 평가자에 의해 실시된 채점 결과에 따라 집단의 수준을 상 · 중 · 하로 나누고, 집단별로 키워드를 산출함으로써 키워드의 상대적 빈도 및 사용 양상에 차이가 있는지를 분석하였다. 이 연구는 키워드가 각 집단별로 일관된 특성을 지닌 차이가 나타날 것을 기대하였으며, 이는 키워드 채점 방식이 쓰기 수행의 수준을 판가름할 수 있는 채점 방식으로서 타당성을 갖는다고 말할 수 있을 것이다.

이 연구는 구체적으로 다음과 같은 연구 질문을 다룬다. 첫째, 각 수준별 집단에 따라 산출된 키워드가 갖는 특성은 무엇인가? 키워드가 채점 방식으로써 타당성을 갖기 위해서는 수험생의 수준을 변별할 수 있어야 한다. 따라서 사전에 평가자에 의해 분류된 수준별 집단에 따라 산출된 키워드가 집단별로 어떠한 특성을 갖는지에 대하여 분석하고자 한다. 이 연구에서는 상위 집단에서 더 의미적이고, 단어 수준이 높으며, 화제에 대한 심층적 논의와 관련된 키워드가 산출될 것이라 기대하였다.

둘째, 상위 집단에서 산출된 키워드와 하위 집단에서 산출된 키워드를 비교하였을 때 어떠한 차이가 있는가? 키워드는 각 집단을 대표하는 성질을 지닌다. 따라서 단순히 키워드가 다른 것에서 그치는 것이 아니라, 키워드가 각 집단에 따른 특성을 가질 것이라는 연구 질문 1의 가설을 바탕으로 같은 지시 대상이나 의미를 가리키는 것일지라도 집단에 따라 서로 다른 단어를 선택할 것이라 기대하였다.

셋째, 키워드 채점 방식이 인간 평가자에 의한 채점에 비하여 타당성을 갖는다고 말할 수 있는가? 키워드 채점은 편의성과 객관성의 측면에서 인간 평가자를 대체할 수 있는 대안적 채점 방식의 하나로 평가받고 있다. 따라서 연구 결과를 바탕으로 키워드 채점 방식이 쓰기 평가에서 활용될 때의 이

점과 한계점을 모색할 수 있을 것으로 기대된다. 특히 수험생의 쓰기 수행을 변별할 수 있는 요인으로써 키워드가 작용할 수 있는지의 여부를 검토함으로써 키워드 채점이 인간 평가자에 의한 채점과 비교하여 타당성을 갖는지의 여부를 검토할 것이다. 이를 통해 점차 교육적 · 사회적 요구가 증대되고 있는 대단위 쓰기 평가의 도입에 필요한 키워드를 활용한 자동 채점 방식을 제안하고자 한다.

II. 연구 방법

1. 연구 대상

이 연구에 참여한 학생들은 충북에 위치한 A 여자 중학교 1학년 학생들(만 13세)이었다. 연구 대상 학교는 군 단위 소재의 학교로 학급당 학생 수는 평균 28.3명이며, 교사 1인당 학생 수는 평균 16.3명이다. 교육 정보 공시 자료에 따르면, 2012학년도에 이 학교의 3학년을 대상으로 실시한 국가수준 학업성취도평가의 국어과 영역에서 보통 학력 이상 86.8%를 기록하였다. 이 학교의 1학년 학생 4개 반 총 113명이 연구에 참여하였다. 국어과 학업성취도 결과 및 지도교사의 견해를 고려하면 위의 학생들은 연구에 참여한 학생들은 일반적인 수준의 중학생 필자를 대표한다고 추정되었다.

또한 주장하는 글에 대해 초등학교 과정에서 학습이 이루어지기는 하지만, 7학년 교육과정에서는 주장하는 글쓰기를 다루지 않기 때문에 학습 효과에 의한 쓰기 수행 편차의 감소가 적어 개인차가 크게 나타나며, 여학생들이므로 쓰기 과제 지시문에 제시된 동물과 관련된 화제에 대한 흥미가 높을 것으로 기대되었다.

2. 검사 도구

1) 쓰기 과제 지시문

쓰기 수행 수준에 따라 나타나는 키워드의 차이를 명확하게 드러내기 위해서는 화제에 따른 학생의 배경지식에 크게 영향을 받지 않으면서도 다채로운 반응을 기대할 수 있는 과제가 요구된다. 따라서 자신의 주장에 따라 근거를 들어 글을 구성하는 '주장하는 글'을 담화 유형으로 선정하였다.

호주 NAPLAN은 초 · 중 · 고 전 학년에 걸쳐 같은 쓰기 과제 지시문을 제공하여 평가한다. 이는 쓰기 과제 텍스트 자체의 해석에서 오는 난이도를 배제하고, 화제에 대한 수험생의 쓰기 수행에 따라 다양하게 나타나는 글의 수준을 변별할 수 있도록 함으로써 학교급과 학년 수준에 관계없이 다양한 스펙트럼의 글을 산출하며, 이에 따라 전 학년에 대한 쓰기 발달 정도를 측정할 수 있다는 이점이 있다. 따라서 NAPLAN Writing 2013에서 제시한 주제인 '동물을 우리에 가둔다고?'라는 주제의 과제를 투입하였다. 이 실험에 쓰인 지시문은 〈표 1〉과 같다.

표 1. 쓰기 과제 지시문

동물을 우리에 가둔다고?
'동물들을 동물원 우리 안에 가두는 것은 잔인하다'는 의견에 대해 어떻게 생각하시나요? 이러한 의견에 대하여 찬성과 반대 중 어느 편을 들어주고 싶은가요? 이에 대한 자신의 생각을 선택하여, 주장하는 글을 20줄(±3줄) 분량으로 써 봅시다(45분).

이 지시문에 제시된 화제는 동물을 동물원 우리 안에 가두어 두는 것에 대한 찬성과 반대 의견을 묻고 있으며, NAPLAN 채점 가이드(NAPLAN 2013 Persuasive Writing Marking Guide)에서 제시한 평가 예시문에 따르면 미숙한 필자의 경우 이 지시문에 대해 단순히 동물의 감정이나 식생, 주거 문제에 치중하며, 능숙한 필자는 생태 환경 및 권리, 자유와 같은 가치에 중점을

두는 것으로 나타났다.

2) 키워드 분석 대상 글

연구 대상 학교의 1학년 총 4개 반을 대상으로 글을 표집하였으며, 총 인원은 여학생 113명이었다. 이 중 불성실하게 작성된 학생의 글 3편이 제외되어, 최종적으로 표집된 글 각 110편에 대한 분석이 이루어졌다.

글 110편을 대상으로 박영민 · 최숙기(2010)에서 사용한 논설문 평가기준표에 따른 총체적 채점을 실시하였다. 국어 교사 및 국어 교육 전공 평가자 3인이 수집된 학생 글 110편을 대상으로 10월 10일부터 10월 25일까지 15일 간 채점을 실시하였다. 채점 결과에 대한 평가자 간 신뢰도를 분석한 결과, Cronbach α 계수는 .833으로 높게 나타났다.

채점 결과에 따라 110편의 글을 상 · 중 · 하 세 집단으로 분류하였다. 상위 집단은 채점 결과 상위 30%에 속하는 학생들로 분류되었으며 글은 총 31편이었다. 하위 집단은 채점 결과 하위 30%에 속하는 학생들을 분류한 것으로 글은 총 33편이었다. 각 집단의 커트라인에서 동점자들이 모두 포함되어 최종적으로 64편의 글이 키워드 분석 대상으로 사용되었다. 〈표 2〉에는 키워드 분석 대상 집단에 대한 언어적 통계량을 제시하였다.

표 2. 분석 대상 글의 언어 통계량

집단	글 수	글자 수 (공백 포함)	글자 수 (공백 제외)	단어 수
상위 집단	31	23,031	17,263	5,742
하위 집단	33	15,008	11,227	3,773
계	64	38,039	28,490	9,515

두 집단의 총 단어수는 9,515개로 집계되었으며, 공백을 제외한 글자수는 28,490개였다. 이 분석 대상에서 중위 집단은 제외되었는데 그 이유는 상위 집단과 하위 집단을 제외한 약 60% 학생들이 표집되므로 표집 대상이 너

무 크고 다양하여 산출된 키워드에서 응집력 있는 특징을 찾기가 어렵고, 상하위 두 집단의 특성을 보다 명시적으로 보여 줄 것을 기대하였기 때문이다.

3. 연구 절차

참여자들은 7학년 학생들로, 4개 반 총 113명의 학생들이 45분 간 제시된 쓰기 과제 지시문을 읽고 글을 작성하였다. 학생들은 이 연구의 취지를 듣고 자발적으로 참여하였으며, 지시문을 읽고 900자 분량의 원고지에 글을 작성하였다. 이에 따라 작성된 글 총 113편이 수집되었으며, 이 중 불성실하게 응답한 3편이 제외되어 글 총 110편에 대한 채점이 실시되었다. 평가자는 국어교사 1인, 2년 이상의 교육 경력을 지닌 국어교육 전공자 2인이었으며, 평가기준표에 따라 채점을 실시하였다. 이 과정에서 연구의 취지 및 채점 시 유의사항에 대한 안내가 제공되었다.

평가자 3인의 채점 결과에 대한 평가자 간 신뢰도의 측정이 실시되었다. 신뢰도는 Cronbach α 계수로 측정되었으며, 평가 결과는 3인 평가자의 내용 · 조직 · 표현 · 단어 선택 · 형식 및 어법의 다섯 가지 평가 요소와 총점의 3인 평가자가 산출한 점수의 평균으로 계산되었다.

그중 총점을 기준으로 각 집단을 상위 30%와 하위 30%로 구분하였으며, 각 집단에 포함된 글에 대한 키워드 산출이 이루어졌다. 30%에 해당하는 동점자를 포함하였으므로, 상위 집단에는 총 31편, 하위 집단에는 총 33편의 글이 표집되었다.

4. 데이터 분석

1) 표집된 글의 형태소 분석

키워드 분석을 위하여 먼저 각 집단에서 제공된 글을 모두 한글 파일로 전사하였다. 전사 과정에서 글에 나타난 맞춤법 오류는 모두 수정되었다. 형

태소 분석기는 맞춤법 오류를 고려하지 않아 형태가 다르면 모두 다른 형태소로 처리하기 때문에, 키워드 산출을 위한 정확한 빈도 계산을 위해서는 형태를 올바른 맞춤법으로 교정하여야 하기 때문이다.[1]

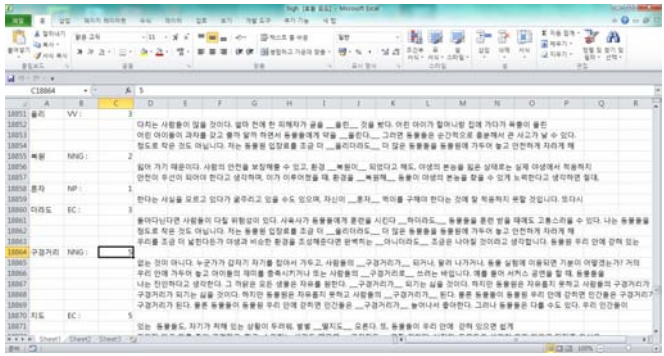

그림 1. 상 집단 학생 글 키워드 빈도 분석 예시

전사된 한글 파일(.hwp)은 텍스트 파일(.txt)로 변환 후 한글 말뭉치 처리 프로그램[2]에 의하여 형태소 단위로 분절하고, 각 형태소의 빈도를 계산하였다. 위의 〈그림 1〉은 한글 말뭉치 처리 프로그램에 의해 분석된 학생 글의 형태소들을 나타낸 것이다.

---

1 실제로 컴퓨터에 의한 키워드 채점이 실시되는 경우, 학생이 정답에 해당하는 키워드를 맞춤법이 틀리게 작성하였을 경우, 오답으로 처리될 것이다. 이 연구는 키워드 채점 방식의 타당성을 분석하고자 하는 연구이므로 맞춤법 오류를 전부 교정하여 전사하였다.

2 일반적으로 영미권 연구에서는 코퍼스 분석을 통한 단어 빈도 산출에 WordSmith corpus analysis package(Scott, 1999) 프로그램을 일반적으로 사용한다. 이 프로그램은 한국어를 지원하지 않는다. 따라서 이 연구에서는 국립국어원 한글박물관에서 제공하는 한글 말뭉치 처리 프로그램(http://www.hangeulmuseum.org/sub/future/information/han_utility01.jsp)을 사용하여 형태소를 분석하였다. 이에 대해서는 홍윤표(2012)에 소개되어 있다.

2) 자동 키워드 산출 과정

키워드 산출은 특정 집단의 언어 자료에서 나타나는 단어의 빈도를 바탕으로 한다. 그러나 단순히 특정 집단에서 어떤 단어의 빈도가 높다고 해서 그 집단의 키워드로 추출되는 것은 아니며, 집단 간 상대빈도로 계산된다. 즉 상대 집단에서 적게 사용된 단어가 해당 집단에서 많이 사용될수록 키워드로써 의미를 갖게 되는 것이다. 이러한 방법으로 키워드를 산출하는 통계적 방법에는 이 연구에서 사용된 방법 외에 카이제곱(x^2), Wilcoxon-Mann-Whitney, 로그우도비(loglikeihood-ratio) 검증 등의 방법이 있다. 이러한 방법들은 각자 장단점을 가지고 있으며 각 키워드 추출 방법에 따라 다른 키워드가 추출될 수 있다.

이 연구에서는 홍종선 외(2001)에서 활용한 t-점수를 활용하여 키워드를 산출하였다.[3] t-점수는 현재 한국어 코퍼스 분석에서 가장 일반적으로 사용되는 방식이다. t-점수가 아닌 다른 측정 방법들은 단어의 빈도가 낮은 경우 측정값이 지나치게 급증한다는 문제가 있으며(김일환 · 이도길, 2011), 또한 t-점수는 상대빈도로 산출되기 때문에 단순 빈도 계산 방식보다 해당 집단의 특성을 더 잘 드러낼 수 있다.

t-점수는 관찰빈도(observed frequency, O)와 기대빈도(expencted frequency, E)의 비교를 통해 상대빈도를 산출하기 때문에 각 집단 크기에 따라 정규화(normalize)된다. 각 집단별 데이터에서 관찰한 빈도를 관찰빈도라 한다. 각 단어들이 해당 집단에서 나타난 빈도를 계산한 유관표(contingency table)는 각 단어에 대한 각 집단별 관찰빈도를 보여 준다. 이 연구에서 키워드 분석을 위하여 단어는 각 형태소별로 처리되었기 때문에 형태

3 이 연구에서는 세 가지 공식을 통해 키워드를 산출하였다. $(O-E)^2/E$(Oaks & Farrow, 2007), $(O-E)^2/\sqrt{E}$(Haberman, 1973), $O-E/\sqrt{O}$(홍종선 외, 2001)을 통해 산출한 결과를 비교하였으며, 앞의 두 공식의 경우 이 연구의 특성상 언어 코퍼스에 비해 표집 규모가 작아 기대빈도가 관측빈도보다 높았던 몇몇 키워드가 음의 값을 지니는 것을 반영하지 못하는 문제가 있어 최종적으로 $O-E/\sqrt{O}$(홍종선 외, 2001)을 채택하였다.

소 단위로 분할된 것들 가운데 형태소이자 단어가 되는 체언 및 용언으로 한정하였으며, 조사는 제외되었다. 아래 〈표 3〉은 가나다순으로 정렬된 단어의 관찰빈도에 대한 세 단어 '가다', '산책', '힘없다'의 사례를 보여 준다.

표 1. 수준별 집단의 단어 관찰빈도

단어	품사	상위 집단	하위 집단	행 합계
가	VV	17	53	70
:	:	:	:	:
산책	NNG	3	0	3
:	:	:	:	:
힘없	VA	1	0	1
열 합계		34,169	32,456	66,625

이 관찰빈도를 바탕으로 하여 기대빈도를 구할 수 있다. 영가설이 참일 때 각 집단별 관찰빈도에 대한 기댓값을 기대빈도라 한다. 예를 들어 상위 집단의 특정 단어에 대한 기대빈도(E)는 다음과 같은 공식에 의해 산출된다. 여기서 $|H|$는 상 집단 글에 포함된 총 단어 수이며, $|L|$은 하위 집단 글에 포함된 총 단어 수를 의미한다. $f_L(\omega)$은 하위 집단(L)에서 단어 w가 나타난 빈도이다.

$$E = \frac{|H| \times f_L(w)}{|L|}$$

기대빈도가 구해지면 이를 바탕으로 각 집단의 단어별 t-점수를 산출할 수 있다. t-점수는 다음과 같은 공식으로 산출된다.

$$t = \frac{O - E}{\sqrt{O}}$$

기대빈도와 t-점수는 각 집단별 관찰빈도 데이터에 Microsoft Excel 2010을 활용하여 수식으로 계산하였다. t-점수는 가장 높은 점수 순으로 정

렬하고, 여기서 실제 채점에서 키워드로 사용될 수 있으며 실질적인 의미를 지닌 품사의 단어만을 대상으로 하기 위하여 〈표 4〉와 같이 일반명사(NNG) · 동사(VV) · 형용사(VA)만을 대상으로 산출하였으며, 나머지 품사의 단어는 키워드 분석 대상에서 제외하였다.

표 4. 품사 분류 기호 및 키워드 분석 대상

대분류	태그	설명	대분류	태그	설명
체언	NNG	일반 명사	어말어미	EF	종결어미
	NNP	고유 명사		EC	연결어미
	NNB	의존 명사		ETN	명사형 전성 어미
	NR	수사		ETM	관형형 전성 어미
	NP	대명사	선어말어미	EP	선어말 어미
용언	VV	동사		EP	
	VA	형용사	접두사	XPN	체언 접두사
	VX	보조 용언	접미사	XSN	명사 파생 접미사
	VCP	긍정 지정사		XSV	동사 파생 접미사
	VCN	부정 지정사		XSA	형용사 파생 접미사
관형사	MM	관형사	어근	XR	어근
부사	MAG	일반 부사	부호	SF	마침표, 물음표, 느낌표
	MAJ	접속 부사		SP	쉼표, 가운뎃점, 콜론, 빗금
감탄사	IC	감탄사		SS	따옴표, 괄호표, 줄표
조사	JKS	주격 조사		SE	줄임표
	JKC	보격 조사		SO	줄임표(물결, 숨김, 빠짐)
	JKG	관형격 조사		SW	기타기호
	JKO	목적격 조사	분석 불능	NF	명사 추정 범주
	JKB	부사격 조사		NV	용언 추정 범주
	JKV	호격 조사		NA	분석 불능 범주
	JKQ	인용격 조사	한글 이외	SL	외국어
	JX	보조사		SH	한자
	JC	접속 조사		SN	숫자

3) 키워드 사용 양상 분석

키워드 사용 양상은 자동적으로 추출된 키워드들이 갖는 경향을 특정 범주로 분류, 분석하여 각 집단별 키워드가 보이는 특성으로 항목화하였다(연구 질문 1). 또한 키워드 사용 양상에서 상위 집단과 하위 집단 간의 비교가 가능한 키워드의 경우, 대응짝을 설정하고 이에 대한 비교 분석을 실시하였다(연구 질문 2).

두 집단 간의 키워드 점수에서 나타나는 차이가 있는지의 여부를 확인하기 위하여 전체 키워드 점수를 대상으로 PASW 18.0의 Wilcoxon 부호 순위 검정을 실시하였다. 이에 따라 각 집단을 변별하는 유의한 변인으로써 키워드가 작용하고 있는지의 여부를 살펴보았다.

마지막으로 키워드가 단순히 단어의 난도와 같은 표층적 수준으로 작용하는지, 아니면 화제 분석이나 논지 전개의 층위에 따라 심층적 수준으로 작용하는지의 여부를 모색해 보고자 윤창욱(2006)에서 제시한 어휘 목록표에 따른 키워드의 난이도를 추정, 집단 간 비교를 실시하였다(연구 질문 3).

III. 연구 결과

참여자들이 작성한 글을 상위 집단과 하위 집단으로 구분하여 t-점수 상위 키워드를 산출하였다. 결측값[4]을 제외한 t-점수는 최소 −288.15점부터 최대 19.62점까지의 범위로 나타났으며, 전체 누적 백분율 42.2%(0.02

4 특정 단어의 관찰빈도가 두 집단 중 어느 하나에서 0인 경우, 해당 집단에서의 t-점수는 계산되지 않는다. 예를 들어 '아저씨(NNG)'의 경우, 상위 집단의 관찰 빈도는 0이고 하위 집단의 관찰 빈도는 7이었는데, 이를 바탕으로 상 집단의 t-점수를 계산하면 $0-10.65/\sqrt{0}$이 되므로 엑셀 수식에서 '#DIV/0!' 오류 메시지가 산출된다. 이러한 값들은 평균 계산에서 결측값으로 처리하여 평균 계산에서 제외되었다.

점)부터 양의 값을 갖는 점수 분포를 보였다(M=-1.67, SD=.35). 기대빈도보다 관찰빈도가 크게 못 미치는 단어들의 경우 높은 음의 값을 갖는 경향이 있다. t-점수가 높을수록 그 단어는 그 집단을 대표하는 키워드로서의 성격을 갖는다.

연구 질문은 (1) 각 수준별 집단에 따른 키워드의 특성, (2) 상 집단에서 산출된 키워드와 하 집단에서 산출된 키워드의 비교, (3) 인간 평가자에 의한 채점에 대한 키워드 채점 방식이 갖는 타당성의 검증에 관심을 두었다. 앞의 두 연구 질문은 t-점수가 높은 순으로 산출된 키워드의 목록에 대한 기술 및 특성에 대한 분류 및 분석을 통해 설명한다. 세 번째 연구 질문은 두 집단에 대한 Wilcoxon 부호 순위 검증 및 단어 난이도의 분석을 통해 설명된다.

1. 집단별 키워드 산출 결과

1) 상위 집단 키워드 산출 결과

t-점수를 바탕으로 하여 각 집단별로 t-점수 50위까지의 키워드를 선정하였다. 먼저 상위 집단에서 산출한 키워드 58개가 〈표 5〉에 제시되었다. t-점수가 높다는 것은 상위 집단에서 하위 집단에 비해 보다 높은 빈도로 사용되었다는 의미이며, 상위 집단의 특성을 드러내는 키워드로써 산출된

표 5. 상위 집단 키워드 목록

인간(4.68), 야생(3.63), 편하다(3.61), 아이(3.32), 감옥(3.32), 속(3.24), 만들다(3.24), 환경(3.17), 입장(3.02), 안전(3.02), 밥(3.00), 다치다(3.00), 조련사(2.83), 육식(2.83), 생각(2.82), 같히다(2.70), 풀다(2.69), 인하다(2.68), 적응(2.65), 기르다(2.65), 감정(2.65), 먹이(2.61), 힘(2.45), 케이지(2.45), 돌고래(2.45), 받다(2.43), 생명(2.40), 일(2.30), 보호(2.30), 어떻다(2.29), 세상(2.29), 본능(2.29), 공간(2.28), 종류(2.24), 장난(2.24), 자라다(2.24), 쇼(2.24), 손길(2.24), 생태계(2.24), 구하다(2.24), 구경거리(2.24), 구경(2.20), 자연(2.11), 잘못(2.07), 최고(2.00), 초식(2.00), 죄(2.00), 조성(2.00), 잡히다(2.00), 야외(2.00), 쉽다(2.00), 삶(2.00), 부모(2.00), 닿다(2.00), 끼치다(2.00), 기회(2.00), 관심(2.00), 건강(2.00)

단어들의 목록이 된다. 상위 집단에서는 50위 커트라인에서 동점(2.00점)에 해당하는 단어들이 모두 포함되어 총 58개의 키워드가 추출되었다.

집단별 키워드 산출 결과를 바탕으로 이를 내용, 조직, 표현의 세 요소로 대분류하여 집단별 특성을 정리하였다. 결과적으로 보면, 내용 면에서 집단별 키워드의 특성이 가장 두드러졌다. 먼저 '내용' 요소에 대한 상위 집단에서 추출된 키워드를 분석한 결과는 다음과 같다.

첫째, 화제에서 논의할 수 있는 가장 구체적이고 심층적인 내용들을 보여 주는 키워드가 나타났다. 이 과제에서는 동물의 생태 환경과 관련된 단어들이 많이 등장한다.

(가) 야생, 감옥, 환경, 밥, 조련사, 육식, 먹이, 케이지, 세상, 공간, 생태계, 자연, 초식, 조성, 야외, 부모, 건강

이는 동물원에서 동물의 생태에 최적화된 환경을 조성하고 가두어 두는 것이 옳은 것인지, 아니면 야생 상태에서 자유롭게 풀어놓는 것이 좋은지에 대한 논의가 주로 상위 집단에서 이루어졌음을 의미한다. 지시문에서 주어진 화제에 대하여 상위 집단의 학생들은 동물의 생태 환경을 고려하여 깊이 있는 논지를 전개하는 경향을 보였다. 또한 (가)의 키워드를 재분류해 보면, 식생활(밥 · 육식 · 먹이 · 초식)과 서식지(야생 · 감옥 · 케이지 · 공간 · 야외), 관계(조련사 · 부모)와 같이 생태 환경 중에서도 매우 다양한 측면을 다루고 있음을 알 수 있다.

둘째, 사안에 대하여 보다 다양한 관점을 선택하여 글을 전개함으로써 화제에 대한 폭넓은 시각을 보여 주는 키워드가 산출되었다. (나)를 보면 동물권(animal right)과 관련된 단어들이 많이 등장한다.

(나) 입장, 안전, 풀다, 적응, 생명, 보호, 본능, 장난, 손길, 구경거리, 구경, 조성, 삶, 관심

'동물권'은 동물을 보호하거나 방생하는 것 중에서 어느 것이 더 가치로운가를 판단하는 데 있어 매우 핵심적인 개념이라 할 수 있는데 상위 집단에서는 이를 다루는 단어들이 매우 많이 사용되었다. 동물의 관점에서 사태를 다루거나(입장 · 적응 · 장난 · 구경거리 · 구경), 동물을 보호하는 관점(안전 · 보호 · 손길 · 조성 · 관심), 동물을 방생하는 관점(풀다 · 본능), 가치와 관련된 관점(생명 · 본능 · 삶)의 단어들이 주로 등장한다. 이는 상위 집단에서 주로 자기 자신뿐만 아니라 여러 가지 관점에서 사안을 바라보고 논의하면서 내용을 전개하였음을 의미한다.

상위 집단에서 추출된 '조직' 요소와 관련된 키워드를 분석한 결과, 인과 관계를 나타내는 단어들이 상위 50위 키워드에 포함되어 있다.

(다) 인(因)하다, 끼치다

상위 집단의 학생들을 일반적으로 글을 쓸 때 '조직'에서 매우 긴밀하고 응집성있게 글을 쓴다. 이에 따라 앞뒤 내용을 긴밀하게 연결하기 위하여 인과 관계를 나타내는 단어를 많이 사용하였으며, 이를 통해 동물을 가두어 두거나 방생하는 경우 일어날 수 있는 가능성과 영향력에 대하여 주로 논의하였다.

2) 하위 집단 키워드 산출 결과

하위 집단의 키워드에는 t-점수 커트라인(1.67점) 동점 단어를 포함하여 총 52개 단어가 포함되었다. 아래 〈표 6〉에 하위 집단에서 t-점수 50위까지에 해당하는 키워드의 목록을 제시하였다.

'내용' 요소에서는 첫째, 쓰기 과제 지시문[5]에서 직접 인용된 단어가 특히 두드러지게 나타난다.

---

5 본고 〈표 1〉 참고.

표 6. 하위 집단 키워드 목록

가두다(4.83), 동물(3.32), 보다(3.01), 개(2.95), 주다(2.92), 키우다(2.90), 안(2.72), 살다(2.69), 아저씨(2.65), 음식(2.60), 싫어하다(2.54), 찬성(2.53), 소(2.5), 불편(2.45), 곰(2.36), 묶다(2.24), 기린(2.24), 병(2.18), 불쌍하다(2.13), 투견(2.00), 주인(2.00), 사막(2.00), 돌보다(2.00), 난폭(2.00), 나타나다(2.00), 교통(2.00), 몸집(1.94), 걸리다(1.94), 같다(1.93), 해치다(1.91), 사고(1.90), 놀다(1.90), 강아지(1.90), 피(1.73), 탈출(1.73), 친구(1.73), 철장(1.73), 존중받다(1.73), 웃다(1.73), 예쁘다(1.73), 여우(1.73), 쓰레기통(1.73), 산책(1.73), 사료(1.73), 물리다(1.73), 독사(1.73), 그릇(1.73), 귀엽다(1.73), 행복(1.67), 태어나다(1.67), 뛰다(1.67), 크다(1.67)

(라) 가두다, 동물, 안(內), 찬성

하위 집단 학생들은 과제 지시문의 단어를 그대로 가져와서 재진술하는 경우가 많았으며, 특히 (라)의 단어들은 대부분 하위 집단의 t-점수 기준 최상위권에 위치하고 있다. 이는 하위 집단 학생들이 글을 쓸 때 과제를 단순히 직접 인용, 반복 진술하는 특성을 보여 준다.

둘째, 내용 전개 방법 중에서도 특히 예시를 많이 사용하고 있음을 드러내는 키워드가 산출되었다. 이는 특정 동물의 명칭이 상 집단에 비해 하위 집단에서 대거 포함된 것을 통해 알 수 있다.[6]

(마) 개, 소, 곰, 기린, 강아지

이는 여러 가지 내용 전개 방법 중에서 자신에게 가장 친숙한 예시를 단

6 이 중 키워드 '강아지'가 추출된 사례들을 제시하면 다음과 같다.
- 거의 새끼 _강아지, _ 새끼 고양이, 토끼 등 귀여운 동물들을 좋아합니다.
- 그런데 __강아지들은__ 열심히 싸워서 피 흘리고 죽습니다.
- 사람들 __강아지에게__ 해준 것이 있을까요?
- 자기들은 __강아지에게__ 때리고
- 러닝머신에 _강아지_ 묶어 놓고 뛰게 하여
- 먹다 남은 이상한 음식, 썩은 음식, 생닭. _강아지_ 시체를 줍니다.
- _강아지도_ 존중받아야 하는 존재니까요.

순히 열거하는 하위 집단의 특성을 드러낸다. 자기 주변에서 일어나거나 겪은 사례를 참고하여 적었으나 심층적인 화제 분석에는 이르지 못하고 주로 자신이 키우는 반려동물과의 경험이나 TV를 통해 겪은 동물 학대와 같은 사례를 중심으로 서술하였다.

'조직' 요소에서, 글의 화제와 연관성이 매우 떨어지는 단어들을 사용함으로써 중심 내용에서 벗어나는 양상을 보여 준다.

(사) 아저씨, 투견, 사막, 피, 친구, 웃다, 예쁘다, 쓰레기통, 산책, 독사, 그릇, 귀엽다

글의 통일성과 응집성은 글의 수준을 결정하는 매우 중요한 요소이다. 그러나 하위 집단의 학생들은 과제에서 제시된 글의 화제와 동떨어진 단어들을 사용하면서 중심 논지를 흐리는 양상을 보인다. 이는 주로 논지 전개 과정에서 사례를 제시할 때 나타나는데 동물과 관련된 사례이기는 하나 동물의 권리나 환경과 같은 심층적인 논의가 아니라 동물로 인한 피해나 학대와 같이 자극적이고 언론에 많이 노출되거나, 또는 반려동물과 관련된 경험과 같은 주로 개인적 수준의 사례 제시에 머무른다는 점에서 하위 집단 학생들이 주변적이고 미시적인 정보에 집중하는 글쓰기 특성을 드러내고 있다.

'표현' 요소에서는 주관성이 높은 감정 관련어가 많이 등장한 점이 특징적이다.

(바) 싫어하다, 불쌍하다, 난폭, 예쁘다, 귀엽다

쓰기 과제 지시문에서 요구하는 문종인 주장하는 글은 글쓴이가 객관적인 관점을 유지하며 논지를 전개하여야 하는 데 반해, 하위 집단 학생들은 필자의 주관성이 개입되는 감정 관련어를 많이 사용하여 글의 객관성을 저해하는 양상을 보인다.

2. 집단별 키워드 대응 비교

상위 집단과 하위 집단에서 나타난 키워드의 상호 비교를 통해서도 각 집단의 특성이 드러난다. 집단별 키워드 간 비교를 통해 나타나는 특성을 정리하면 다음과 같다.

첫째, 입장을 정하고 진술할 때에 필자 자신의 관점을 객관화하는 정도에서 차이가 나타난다. 키워드 (아)와 (자)를 보면 상위 집단에서는 주로 동물의 관점을, 하위 집단은 주로 인간의 관점을 대변하는 키워드가 사용된다.

(아) 갇히다: 가두다
잡히다: 키우다

(자) 편하다: 불편
안전: 교통 · 사고 · 물리다

일반적으로 능숙한 필자는 글을 쓸 때 예상 독자를 고려하며 다양한 관점을 취하여 글을 쓴다. 상위 집단의 학생들은 자신에게는 타자인 동물의 관점에서 서술하는 경향이 있으며, 이는 특히 피동사를 사용한 키워드(갇히다 · 잡히다)를 통해 나타났다. 즉 주어로 동물을 상정하고 그에 따라 피동사를 사용하는 양상이 하위 집단에 비해 두드러졌다는 의미이다. 또한 동물의 입장에서 동물의 안위와 관련된 단어(편하다 · 안전)를 사용하였다. 이는 타자의 입장에서 사태를 객관적으로 바라보는 상위 집단 학생들의 특성을 나타낸다.

반면 하위 집단의 학생들은 자신과 같은 인간의 관점에서 서술함으로써 사동사(가두다 · 키우다)가 주로 나타났다. 또한 인간의 입장에서 동물을 방생하였을 때 나타날 문제점들을 보여 주는 단어들(불편 · 교통 · 사고 · 물리다)을 사용하였다. 이는 하위 집단에서 필자 중심적 글쓰기(writer-centric

writing)가 나타남을 의미한다. 초보적인 필자는 예상독자의 요구나 기대, 관점, 태도, 배경지식 등을 고려하지 않고 필자인 자기 자신의 관점에서 독백하듯 글을 쓰는 자기 중심적인 관점을 나타내는 경향이 있는데(박영민, 2005) 하위 집단의 키워드는 독자의 관점을 고려하지 않고 주로 자신의 입장에서 주장을 펼치는 전형적인 하위 집단 글쓰기의 특성을 드러낸다. 즉 하위 집단의 학생들은 화제에 대한 다양하고 객관적인 시각보다는 주로 자신의 관점에서 입장을 취하는 경향이 키워드를 통해 나타난 것이다.

둘째, 같은 의미를 지니는 단어라 할지라도 두 집단에서 사용하는 단어가 다른 경우가 있다. 상위 집단 학생들은 중립적이고 공식적인 단어를 사용하는 반면, 하위 집단 학생들은 편향적이고 비공식적인 구어체의 단어를 사용하는 경향이 나타났다.

(차) 케이지: 철장
먹이: 음식 · 사료

먼저 '동물을 가두어 기르는 곳'의 의미를 지니는 단어가 집단별로 다르게 쓰이고 있음이 나타났다. 상위 집단에서는 케이지라는 외래어를 사용하였고 하위 집단에서는 '철장'이라는 한자어를 사용하였는데, 여기서 두 단어는 같은 지시 대상을 가리키는 단어이지만 의미역이 서로 다른 것이 특징적이다. '케이지'가 중립적인 의미역을 가짐에 반하여 '철장'은 대체로 부정적 의미가 강하게 나타나는 단어이다. 이는 상위 집단 학생들이 중립적인 단어를 선택하여 글을 쓰는 반면, 하위 집단 학생들은 다소 편향적인 의미역을 지닌 단어를 선택하여 글을 쓴다는 것을 나타낸다.

또한 상위 집단에서 사용한 '먹이'라는 단어는 주로 문어체에서 주로 나타나는 공식성이 강한 단어임에 반하여, '음식 · 사료'는 일상적인 구어체에서도 많이 사용되는 비공식성이 강한 단어라는 점에서 차이점이 있다. 상위 집단의 학생들은 담화공동체의 문어 양식에 익숙하여 단어를 선택할 때에

문어 양식에 가까운 단어를 선택하여 사용하지만, 하위 집단의 학생들은 평소에 자신이 사용하는 단어를 글을 쓸 때에도 일관되게 적용하였다.

글에서 어느 한 쪽에 치우치지 않은 중립적인 단어를 선택하여 사용하는 것은 주장하는 글쓰기에서 객관성을 유지하기 위하여 요구되는 필수적 요소라 할 수 있다. 또한 담화 공동체의 문체 양식을 익히고 문종과 예상독자를 고려하여 글을 쓰는 것은 쓰기 수행에서 능숙한 필자가 보이는 특징 중 하나이다. 즉 상위 집단의 학생들은 키워드를 통해 독자 중심적 글쓰기(audience-centric writing) 경향을 보여 주고 있다.

키워드 간 비교 결과, 상위 집단은 공식적이고 중립적인 단어를 선택하여 다양한 관점에서 필자를 객관화하고 예상독자를 고려하여 글을 쓰는 독자 중심적 글쓰기를 보여 주는 반면, 하위 집단은 비공식적이고 편향적 단어를 선택하여 주로 자신의 입장에서 논지를 전개하는 필자 중심적 글쓰기가 나타난다는 것을 보여 주었다.

3. 키워드 채점 방식의 타당성

키워드 채점 방식은 화제에 대한 수험생의 사고의 깊이와 쓰기 과제에 대한 분석 능력을 판정할 수 있는 채점 방식이다. 인간 평가자의 채점 결과로 얻어진 상 · 하위 집단 구분에서 추출된 키워드 전체를 대상으로 점수에 대한 Wilcoxon 부호 순위 검정[7]을 시행한 결과, 두 집단의 키워드에 대한 점수를 비교하였을 때 〈표 7〉과 같이 통계적으로 유의미한 차이가 나타났다 ($z=-4.231$, $p=.000$).

---

7 F. Wilcoxon에 의해 제안된 비모수 통계검정방법으로, 종속표본 t검증이 서로 독립적이지 않은 두 집단의 집단 간 차이를 검증하는 모수통계검정방법이라면, Wilcoxon 검정은 이에 대응하는 비모수 통계검정방법이다.

표 7. 키워드 t-점수에 대한 Wilcoxon 부호순위 검정 결과

	N	평균순위	순위합	
t-점수(하위 집단) 〈 t-점수(상위 집단)	282	342.03	96,453.00	z = -4.231 p = .000** z = -4.231 p = .000**

(음의 순위 기준)

이는 키워드가 상위 집단과 하위 집단에서 서로 다른 양상으로 나타났던 III-1과 III-2의 연구 결과를 지지하는 것이다. 따라서 키워드를 통해 쓰기 수행을 평가하는 방식은 인간 평가자와 동일하게 쓰기 수행이 능숙한 필자와 미숙한 필자를 변별하였다. 즉, 키워드 채점 방식에 의해 수험생의 쓰기 수행을 채점하는 것은 타당하다고 결론지을 수 있다.

다만 키워드가 단순히 단어 수준의 채점 방식이라 하여 이를 단어의 난이도에 따른 채점으로 오해하여서는 안 된다. 분명하게 키워드 채점 방식은 단순히 단어의 난이도로 측정되는 것이 아니다. 이는 이독성(readability) 공식 산출을 위한 윤창욱(2006)의 어휘 목록에 따른 어려운 단어 판정을 통해서도 확인할 수 있다. 윤창욱(2006)의 전체 5,178단어로 이루어진 어휘 목록표에 실려 있지 않은 단어는 어려운 단어로 간주된다. 상위 집단과 하위 집단의 상위 50위 키워드에 대한 단어의 난이도 수준을 분석한 결과는 다음 〈표 8〉과 같다.

표 8. 윤창욱(2006) 어휘 목록에 따른 집단별 단어 난이도 분석

	상위 집단	하위 집단	전체
쉬운 단어	50(86.2%)	40(69.0%)	90(81.82%)
어려운 단어	8(13.8%)	12(20.7%)	20(18.18%)
계	58(100%)	52(100%)	110(100%)

단어의 난이도 분석 결과 상위 집단의 어려운 단어 비율인 13.8%(n=8)보다 오히려 하위 집단의 단어에서 어려운 단어의 비율이 20.7%(n=12)로 더 높

게 나타났다. 이를 구체적으로 예시하면 위의 〈표 9〉와 같다. 집단별 상위 50개 키워드 중 하위 집단에 어려운 단어가 더 많이 포함되었음을 알 수 있다.

표 9. 윤창욱(2006)에서 어려운 단어로 분류된 집단별 키워드

상위 집단	하위 집단
조련사(2.83), 적응(2.65), 케이지(2.45), 돌고래(2.45), 구경거리(2.24), 조성(2.00), 야외(2.00)	가두다(4.83), 기린(2.24), 투견(2.00), 난폭(2.00), 해치다(1.91), 탈출(1.73), 철장(1.73), 존중받다(1.73), 여우(1.73), 산책(1.73), 사료(1.73), 독사(1.73)

따라서 키워드 채점 방식은 단순히 단어의 난이도로 측정되는 것만은 아님을 알 수 있다. 어려운 단어를 많이 쓴다고 해서 글의 수준이 올라가는 것이 아님은 당연하며, 이는 이 연구의 결과와도 일치한다. 키워드 채점 방식에서 집단 간 차이를 보인 키워드의 양상이 화제에 대한 분석 능력과 사고의 수준을 반영하는 단어들임에 주목하여야 한다. 키워드 분석 결과 상위 집단의 학생들은 동일한 화제에 대하여 더 심층적이고 분석적인 논의를 이끌어 내면서 여러 가지 관점에서 사안을 바라보고 논의하였다. 또한 공식적이고 중립적인 단어를 선택하여 객관적인 관점을 유지하며 글을 쓰는 경향이 나타났다. 반면 하위 집단의 학생들은 화제에 대한 깊이 있는 분석보다는 주로 자신의 경험이나 언론에서 자주 노출되는 사례들을 근거로 하여 피상적으로 논지를 전개하였으며, 과제 지시문의 단어를 그대로 인용하거나 비공식적이고 편향적인 단어들을 주로 사용하여 글을 쓰는 경향이 있었다.

IV. 결론

이 연구에서는 인간 평가자에 의해 분류된 두 집단에서 키워드를 산출

하여 상위 집단과 하위 집단 간의 키워드 사용 양상이 서로 다르게 사용됨을 밝혔으며, 이를 통해 키워드 채점 방식이 갖는 쓰기 평가 채점 방식으로서의 타당성을 검토하였다.

연구 결과 첫째, 키워드 채점에서 상위 집단의 학생들은 보다 심층적인 화제 분석과 다양한 논지 전개 방식을 활용하여 내용적으로 풍부한 글을 썼으며, 논의 과정에서 사안에 대한 다양한 관점을 다루면서 객관적으로 서술하였다. 조직에 있어서도 인과 관계를 나타내는 단어를 사용하여 보다 긴밀하고 응집성 있는 글을 전개하였다. 또한 표현에 있어서도 주장하는 글이라는 문종과 예상독자, 담화 공동체의 문어 양식을 고려하여 공식적이고 중립적인 단어를 사용하는 독자 중심적 글쓰기의 양상을 보였다. 반면 하위 집단의 학생들은 내용 면에서 과제 지시문의 진술을 그대로 가져와 재진술하는 빈도가 매우 높게 나타났으며, 내용 전개 시에 자신의 경험에 가까운 예시를 주로 사용하는 모습을 보였다. 또한 조직 면에서는 통일성과 응집성이 떨어지는 단어들이 대거 사용되었으며, 표현 면에서는 감정적인 단어들을 많이 사용하여 글의 객관성을 저해하고 일상에서 주로 쓰는 구어체 표현을 사용함으로써 비공식적, 편향적인 성격이 강하게 드러나는 글을 작성함으로써 필자 중심적 글쓰기의 경향이 나타났다.

이는 키워드가 수험생의 수준을 변별하는 요인으로써 효과가 있음을 확인하는 결과이며, 특히 키워드가 단순히 단어의 난이도와 같은 표층적이고 미시적인 수준의 측정이 아니라, 화제에 대한 분석과 논지 전개 방식, 내용의 풍부성, 단어 선택과 같은 평가 요소 전반에 걸쳐 있는 채점 방식임을 의미한다. 따라서 사전에 평가자들이 어떠한 키워드를 설정하느냐에 따라 키워드 채점 방식은 기존의 평가기준표에서 일반적으로 사용되는 5개 평가 요소인 내용 · 조직 · 표현 · 단어 선택 · 형식 및 어법을 아우를 수 있는 채점 방식인 것이다. 이는 단어를 기반으로 하여 컴퓨터 자동 채점이 이루어지는 영미권의 쓰기 평가에서 컴퓨터 자동 채점이 인간 평가자의 점수에 대한 예측력을 연구한 Burstein *et al.*(2001)에서 e-rater가 인간 평가자의 점수를

87~94% 예측하였다는 것에서도 확인할 수 있는 결과이다.

둘째, 키워드 채점은 국어과 평가에서 고등 사고 능력과 창의적 사고의 평가가 요구되는 현 시점에서 대단위 국어과 평가에 직접 쓰기 평가가 도입되기 위하여 필요한 대안적 채점 방식이다. 현재 대단위 쓰기 평가가 시행되지 못하고 있는 가장 큰 원인으로 지적되고 있는 절차적 비용의 문제를 해소할 수 있으며, 또한 쓰기 평가에서 고질적인 문제인 평가자의 주관성 개입을 배제할 수 있다. 따라서 기존의 인간 평가자에 의한 채점의 문제점을 해소하면서도, 현재의 컴퓨터 자동 채점 방식이 갖는 불완전성을 보완하는 가교적 역할을 할 수 있는 채점 방식이다.

현재 대단위 쓰기 평가의 도입을 위한 컴퓨터 자동 채점에 대한 연구(노은희 외, 2012)가 이루어지고 있는 상태에서 키워드에 의한 채점은 단어를 기반으로 한 컴퓨터 자동 채점에서 보다 타당하고 정련된 채점을 할 수 있도록 기초적이고 실증적인 데이터를 제공할 수 있다. 키워드를 통한 채점 방식이 일반화되고 이에 대한 평가 절차가 정선되어 평가 결과가 누적되면 키워드와 평가 결과에 대한 분석을 통해 데이터를 구축하고, 어떤 키워드가 수험생의 쓰기 능력을 더 정확하게 예측할 수 있는지, 더 신뢰할 수 있는 평가 결과를 산출할 수 있는지의 여부를 판단할 수 있는 데이터를 제공할 수 있다. 또한 이러한 데이터를 지속적으로 축적함으로써 코퍼스 수준의 데이터를 구축하면 자연어 기반의 컴퓨터 자동 채점이 보다 더 신뢰롭고 타당한 채점을 지향할 수 있다.

셋째, 키워드 채점은 글의 수준에 대한 양적 측정이 가능하여 평가 결과의 처리와 해석에 수월성을 갖는다. 영미권의 작문 연구자들은 이전부터 글의 수준을 양적으로 측정하는 것에 대한 관심을 기울여왔다(Hunt, 1965; Mellon, 1969; Ellen & Davis, 1980; Gillam & Johnston, 1992; Ukrainetz, 2006). 글의 수준에 대한 양적 측정을 통해 결과의 통계적 처리뿐만 아니라 데이터의 처리를 통하여 글에 대한 여러 가지 분석이 가능하기 때문이다. 언어기반평가(Language-based Assessment)는 이러한 연구의 성과라 할 수

있다. 그중 가장 보편적인 언어기반평가로 알려진 T-Unit은 Hunt(1965)에 의해 제안된 쓰기 평가에서 양적으로 문장의 성숙도(syntactic maturity)를 측정하기 위해 개발된 방식이다. 우리나라에서도 정희모 · 김성희(2008)의 연구에서 대학생 필자의 글을 t-unit 방식으로 채점하고자 하는 시도가 이루어졌다. 그러나 언어기반평가는 단순히 화제에 대한 분석 능력이나 과제 수행과 같은 심층적인 수준의 글의 요소에 대한 채점이 어렵다는 단점이 있다. 언어기반평가는 영어에서 단어 길이, 복합절, 문장 길이와 같은 글의 양적 요인을 통해 글의 수준을 판단하는 데 사용되어 왔다. 영어와 달리 한국어의 언어 구조는 단어 길이, 복합절의 사용이나 문장 길이와 같은 요소가 글의 수준을 판가름한다고 보기가 어렵다. 국어의 이독성에 대한 연구들에서 글의 이독성을 결정하는 요인으로 주로 단어 길이, 문장의 길이, 복합절의 사용 등의 여러 가지 양적 요인을 설정하고는 있으나, 음절 수나 어절 수와 같은 글자 수나 길이에 기반을 둔 양적 요인이 그다지 설명력을 갖지 못했다는 연구 결과(최숙기, 2012)는 이를 뒷받침한다. 반면 키워드 채점 방식은 언어기반평가가 갖는 그간의 글 수준에 대한 양적 측정에서 나타나는 획일성과 단편성에서 벗어나면서도, 보다 인간 평가자 지향적인 채점이 이루어질 수 있다. 키워드 채점 방식은 사전에 평가자 간 협의를 통하여 과제 지시문 분석과 글의 수준에 따른 키워드 선정을 통해 평가 목적, 평가 상황과 같은 평가의 여러 국면을 고려하고 이를 채점에 반영할 수 있으며, 이후 객관적이고 신속한 채점이 이루어질 수 있다는 점에서 언어기반평가가 갖는 한계를 넘어설 수 있다. 또한 채점 결과로 산출된 데이터에 대하여 단어의 빈도나 연어 배열과 같은 단어 수준에서부터 단어 간 의미망 분석과 같은 다층적 분석 방식에 이르기까지 다양한 분석 방법의 적용을 통해 평가 결과의 피드백에 있어서도 보다 많은 정보를 제공할 수 있다는 이점이 있다.

이 연구에서는 키워드 산출을 위하여 단어의 상대 빈도를 바탕으로 산출되는 t-점수를 적용하였다. 집단 간 키워드의 차이를 명시적으로 드러내기 위하여 이 연구에서는 중위 집단을 제거하고 키워드를 산출하였다. 즉, 이

연구의 한계점은 중위 집단에 대한 데이터 분석이 이루어지지 않았다는 점이다. 양 극단의 데이터 분석이 보여 주는 명확함에도 불구하고, 이는 전체 집단 중 약 60%의 수험생들이 위치한 중위 집단의 데이터에 대한 결과를 보여 주지 못했다. 따라서 이후 중위 집단을 포함한 키워드 산출을 통해 보다 면밀하게 수험생의 쓰기 수행과 키워드 간의 상관관계를 연구할 필요가 있다.

또한 이 연구의 대상은 7학년 학생이기 때문에 키워드 채점과 같은 쓰기 평가가 주로 치러지는 고등 수준의 쓰기 평가에까지 키워드 평가의 타당성을 일반화하여 적용하기 위해서는 성인을 포함한 더 다양한 연구 집단을 대상으로 하여 연구가 수행되어야 할 것이다.

한편 최근 화제와 단어들 간의 관계에 대한 다층적인 의미망 분석(박지영 외, 2013; 김민호 · 권혁철, 2011; 한관종, 2003)이 이루어지고 있는 시점에서, 이 연구는 키워드와 키워드, 키워드와 화제 간에 의미적으로 연결될 수 있는 지점이 분명히 관찰되었음에도 불구하고 추가 분석을 실행하지 않았다. 따라서 추후 연구를 통해 키워드 간, 화제와 키워드 간의 의미망 연결을 도식화하여 보여 준다면, 키워드가 평가 결과에 미치는 영향력을 보다 세부적인 수준에서 분석할 수 있을 것이다.

위와 같은 연구의 한계점에도 불구하고, 이 연구는 키워드 채점 방식이 인간 평가자의 채점 결과에 따른 집단 구분에 따라 서로 다른 키워드가 산출되었으며, 키워드에서 나타나는 경향성에 있어서도 차이가 있음을 밝힘으로써 키워드 채점 방식이 갖는 타당성을 검증하였다는 데 의의가 있다. 이 연구가 대단위 쓰기 평가가 도입되고 일반화되는 과정에서 채점 방식에 대한 다면적인 검토 과정에 유용하게 활용될 수 있기를 바란다.

* 본 논문은 2014. 5. 2. 투고되었으며, 2014. 5. 5. 심사가 시작되어 2014. 5. 24. 심사가 종료되었음.

참고문헌

김민호 · 권혁철(2011),「한국어 어휘의미망의 의미 관계를 이용한 어의 중의성 해소」, 『정보과학회논문지: 소프트웨어 및 응용』, 38(10), 554-564.

김일환 · 이도길(2011),「대규모 신문 기사의 자동 키워드 추출과 분석—t-점수를 이용하여」, 『한국어학』 53(단일호), 145-194.

노은희 · 심재호 · 김명화 · 김재훈(2012),「대규모 평가를 위한 서답형 문항 자동채점 방안 연구」, 한국교육과정평가원, 연구보고 RRE2012-6.

박영민(2005),「학생 작문의 다니엘 효과와 예상독자 인식의 방법」, 『새국어교육』 70, pp. 73-99.

박영민 · 최숙기(2010),「중학생 논설문 평가의 모평균 추정과 평가 예시문 선정」, 『국어교육』 21, pp. 69-91.

박지영 · 김태호 · 박한우(2013),「의미연결망 분석을 통한 셀러브리티의 SNS 메시지 탐구」, 『방송통신연구』, 36-74.

윤창욱(2006),「비문학 지문 이독성 공식 개발에 관한 연구」, 학위논문(석사), 한국교원대학교 교육대학원.

정희모 · 김성희(2008),「대학생 글쓰기의 텍스트 비교 분석 연구—능숙한 필자와 미숙한 필자의 텍스트에 나타난 특징을 중심으로」, 『국어교육학연구』 32(단일호), 393-426.

최숙기(2012),「텍스트 복잡도 기반의 읽기 교육용 제재의 정합성 평가 모형 개발 연구」, 『국어교육』 139, 451-490.

한관종(2003),「사회과학 방법론으로서의 연결망 분석기법 적용의 의의와 연구과제—의미와 연결망 분석(semantic network analysis)을 중심으로」, 『사회과교육연구』 10(2), 219-235.

홍윤표(2012), 『국어정보학』, 태학사.

Attali, Y. & Burstein, J.(2006), *Automated essay scoring with e-rater® V. 2. The Journal of Technology, Learning and Assessment*, 4(3).

Burstein, J., Kukich, K., Wolff, S., Lu, J., & Chodorow, M.(2001), *Enriching automated essay scoring using discourse marking*, ERIC Clearinghouse.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D.(1991), "Quality control in the development and use of performance assessments," *Applied measurement in education*, 4(4), 289-303.

Eells, W. C.(1930), "Reliability of repeated grading of essay type examinations," *Journal of Educational Psychology*, 21(1), 48.

Ellen W. Nold & Brent E. Davis, "The Discourse Matrix," *College Composition and Communication* 31(1980), pp. 141-152.

Gillam, R. B., & Johnston, J. R.(1992), "Spoken and written language relationships in language/learning-impaired and normally achieving school-age children," *Journal of Speech, Language, and Hearing Research* 35(6), 1303-1315.

Hughes, D. C., & Keeling, B.(1984), "The use of model essays to reduce context effects in essay scoring," *Journal of Educational Measurement* 21(3), 277-281.

Hyland, Ken, & Polly Tse., "Is there an 'academic vocabulary'?" *TESOL quarterly* 41.2

(2007): 235-253.
Kellogg W. Hunt(1965), *Grammatical Structures Written at The Three Grade levels*, Champaign, II: National Council of Teachers of English.
Kilgarriff, A.(2001), *Comparing corpora*, International journal of corpus linguistics, 6(1), 97-133.
Landauer, T. K., Laham, D., & Foltz, P. W.(2003), "Automated scoring and annotation of essays with the Intelligent Essay Assessor," *Automated essay scoring: A cross-disciplinary perspective*, 87-112.
Landy, F. J., & Farr, J. L.(1983), *The measurement of work performance: Methods, theory, and applications*, New York: Academic Press.
Leech, G., Rayson, P., & Wilson, A.(2001), *Word frequencies in written and spoken English: based on the British National Corpus*, Longman.
Mellon, J. C.(1969), *Transformational sentence-combining: A method for enhancing the development of syntactic fluency in English composition* (No. 10-13), Urbana, IL: National Council of Teachers of English.
Oakes, M. P., & Farrow, M.(2007), "Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries," *Literary and Linguistic Computing* 22(1), 85-99.
Paquot, M., & Bestgen, Y.(2009), "Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction," *Language and Computers* 68(1), 247-269.
Saal, F. E., Downey, R. G., & Lahey, M. A.(1980), "Rating the ratings: Assessing the psychometric quality of rating data," *Psychological Bulletin* 88(2), 413.
Thorndike, E. L.(1920), "A constant error in psychological ratings," *Journal of applied psychology* 4(1), 25-29.
Ukrainetz, T. A.(2006), *Contextualized language intervention*, Eau Claire, WI: Thinking Publications.
Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., and Nevill-Manning, C. G.(1999), "KEA: Practical automatic keyphrase extraction," In *Proceedings of the fourth ACM conference on Digital libraries*, p. 255, ACM.

초록

쓰기 평가의 키워드 채점 방식에 대한 타당성 분석

이지원

이 연구는 키워드 채점 방식이 쓰기 평가에서 수험생의 수준을 변별할 수 있는지를 밝힘으로써, 그 타당성을 분석하고자 하였다. 인간 평가자의 채점을 통해 구분된 상위 집단과 하위 집단의 글에서 t-점수를 이용하여 키워드를 산출한 결과, 두 집단의 키워드에서 나타나는 양상이 매우 다른 것으로 나타나 키워드가 집단을 변별할 수 있는 요소로 작용함을 확인하였다.

연구 결과 상위 집단의 키워드는 보다 심층적인 화제 분석과 다양한 논지 전개 방식을 활용하는 것으로 나타나는 단어들을 사용하였으며, 논의 과정에서 사안에 대한 다양한 관점을 다루면서 객관적으로 서술하였다. 또한 조직에 있어서도 인과 관계를 나타내는 단어를 사용함으로써 보다 응집성 있는 글을 전개하였으며, 단어 선택에서 공식적이며 중립적인 의미를 가지는 단어들을 주로 사용하는 것으로 나타났다. 반면, 하위 집단은 자신의 경험이나 언론 상에서 자주 노출되는 자극적인 사건들을 중심으로 예시를 많이 사용하였으며, 감정 관련어를 사용하여 글의 객관성을 저해하였다. 또한 과제 지시문의 진술을 그대로 가져와 재진술하는 경향이 나타났으며, 단어 선택에서는 필자 중심적인 단어를 선택하여 구어체적이고, 비공식적이며, 편향적인 단어들이 주로 사용되었다.

이는 키워드가 두 집단 간에 서로 다르게 추출되었음을 의미하며, 또한 그 키워드가 서로 다른 경향성을 가짐으로써 집단을 변별하는 데 있어 키워드가 유효하였음을 확인하였다는 데 의의가 있다.

핵심어 키워드, 쓰기 평가, 키워드 채점, 평가방법, 평가자, 자동채점 프로그램, 키워드 산출, 평가자 신뢰도, 필자 중심적 글쓰기, 독자 중심적 글쓰기

ABSTRACT

The Validity on the Keyword Matching Scoring Methods in Writing Assessment

Lee, Ji-won

The purpose of this study was to investigate the validity on keyword matching scoring methods. In keyword extraction results, vocabulary differed according to a variety of writing performance of group classified from holistic rating results by human raters. Such are suggested that keyword matching scoring methods are valid to distinction of writing performance.

As a result, high-level group keywords are deeper insight of topic and used to various ways of development in discussion. They also mainly used to vocabulary indicated the causal nexus in organization, formal and non-biased vocabulary in word selection. On the other side, low-level group keywords are mainly used to enumerate bits of examples way of development in discussion, emotional vocabulary which disrupt objectivity in the text. In addition, they are used to writer-centrism vocabulary which have colloquial, informal, biased traits.

These result indicated that extract distinguished keywords between high-level and low-level group. Therefore, keyword scoring methods have validity in writing assessment system.

KEYWORDS Keyword, Writing Assessment, Assessment Methods, Automatic Keyword Extraction, Rater, Rater Reliability, Automatic Scoring System, writer-centric writing, audience-centric writing