교육과정평가연구
The Journal of Curriculum and Evaluation
2014, Vol. 17, No. 2, pp. 71~97

대규모 영어 단문형 쓰기 평가를 위한 자동채점 프로그램의 적용 가능성 탐색[1)]

시 기 자(한국교육과정평가원 연구위원)*
박 도 영(한국교육과정평가원 부연구위원)
임 황 규(한국교육과정평가원 전문연구원)

≪ 요 약 ≫

본 연구의 목적은 단문 형식의 쓰기 답안 평가를 위해 개발한 자동채점 프로그램의 성능을 검증하여 대규모 쓰기 평가에서의 적용 가능성을 탐색하기 위한 것이다. 본 연구에서 성능 검증에 사용한 문항은 '상황에 맞는 짧은 글쓰기'(15~25단어, 5분) 문항과 '그림의 세부묘사 완성하기'(하위문장별 10단어 이내, 5분) 문항이다. 본 연구에서는 단문 형식의 쓰기 평가를 위한 자동채점 프로그램의 성능을 검증하기 위해 인간채점과 자동채점에 따른 상관계수, 유사일치도 통계에 근거한 채점자 간 신뢰도의 차이, 다국면 라쉬 모형에 근거한 채점자 엄격성의 차이, 일반화가능도 계수에 근거한 검사점수 신뢰도의 차이, 시간 및 비용 차이 등에 대한 통계적 분석을 실시하였다. 분석 결과, 자동채점이 인간채점자 1명을 대체할 경우 채점자 간 신뢰도, 검사점수 신뢰도를 인간채점과 유사한 수준으로 유지하면서 채점자 엄격성에 의한 영향력과 시간 및 비용을 큰 폭으로 감소시킬 수 있음을 확인하였다.

주제어: 단문형 쓰기 평가, 자동채점, 채점자 간 신뢰도, 채점자 엄격성, 일반화가능도 계수

1) 본 연구는 한국교육과정평가원(2013)에서 수행한 '국가영어능력평가시험 쓰기 자동채점 프로그램 개발 (Ⅱ)'의 일부 내용을 발췌・요약한 것임.

* 제1저자 및 교신저자, skj@kice.re.kr

Ⅰ. 서론

지구촌의 세계화(globalization)로 영어가 국가 경쟁력에 중요한 영향을 미치게 됨에 따라 정부에서는 의사소통 중심의 말하기·쓰기 교육을 강화하고 있다. 이와 같은 의사소통 중심의 교육이 제대로 정착되기 위해서는 말하기·쓰기 평가의 타당성과 신뢰성 확보가 중요하게 고려되어야 한다. 현재 대부분의 중등학교에서는 영어 쓰기에 대한 수행평가 문항을 서술형이나 논술형으로 출제하고 있으며 서술형·논술형을 사용하는 비율도 점차 늘어나고 있다. 그러나 단위 학교 또는 국가수준의 대규모 평가에서 말하기, 쓰기 평가를 포함할 경우 채점자 훈련, 채점 소요 시간, 예산, 보안 등 현실적으로 고려해야 할 사항들이 많다.

이러한 말하기, 쓰기 평가 채점의 현실적인 어려움을 해결하기 위해 한국교육과정평가원에서는 2012년부터 국내 기업과 공동으로 대규모 말하기, 쓰기 평가에 적용 가능한 자동채점 프로그램 개발 연구를 수행하고 있다. 2012년에는 에세이 형식의 쓰기 2급 답안(최대 120단어) 채점에 적용하기 위한 자동채점 프로그램을 개발하였으며, 2013년에는 에세이형 자동채점 프로그램을 고도화하고 쓰기 3급 단문 형식의 답안(최대 25단어) 채점에 적용하기 위한 자동채점 프로그램을 개발하였다.

에세이형 자동채점 프로그램에 대한 타당성 검증 결과, 채점자 간 신뢰도는 인간채점과 유사한 수준을 나타냈으며 검사점수의 신뢰도는 0.95 이상의 높은 수준을 유지하였고 채점자 엄격성에 따른 차이가 채점 결과에 미치는 영향력이 크지 않아 자동채점의 적용가능성에 대한 긍정적인 결과가 도출되었다. 또한 프로그램의 성능 향상을 위한 개선 사항으로는 정확한 기준점수를 자료로 활용한 기계학습, 자동채점 예외 답안의 선별 및 처리, 한국 학생들의 언어적 특성을 고려한 대규모 코퍼스 구축, 구문 분석과 관련된 채점자질의 개발, 채점 영역별 채점자질의 보완, 오류 분석 기능의 개선이 필요한 것으로 확인되었다(시기자 외 2013a).

단문 형식의 쓰기 평가 문항은 에세이 형식의 쓰기 평가 문항과 평가 요소가 다르기 때문에 기개발된 자동채점 프로그램을 적용하기에는 한계가 있다. 구체적으로 살펴보면, 에세이 형식의 답안은 답안 텍스트의 응집성(coherence)이나 담화 구조(discourse structure)가 평가 대상에 포함되지만, 단문 형식의 답안에는 이와 같은 요소가 존재하지 않는다. 또한, 에세이 형식의 평가 문항은 특정 주제에 대한 학생들의 경험이나 의견을 묻는 경우가 많은 반면, 단문 형식의 평가 문항은 답안이 정해져 있는 경우가 대부분이다. 따라서 에세이 형식의 평가 답안과 단문 형식의 평가 답안은 채점을 위한 지식 기반(knowledge base)의 규모에 있어 차이가 있다. 단문 형식의 답안을 자동채점하려면 답안 내에 '정답 개념'이 포함되어 있는지의 여부를 판단해야 하며, '정답 개념'에 대한 동의어, 유사어, 어형의 변화, 구문의 다양성, 지시 대명사의 사용 등을 함께 처리할 수 있어야 한다. 한국교육과정평가원에서는 기개발 에세이형 자동채점 프로그램

의 타당성 검증 결과에서 도출된 시사점과 단문 형식의 답안 특성을 고려하여 대규모 단문 형식의 쓰기 답안 채점에 적용 가능한 자동채점 프로그램을 개발하였다(시기자 외, 2013b).

이에 본 연구에서는 에세이 자동채점 프로그램의 타당성 검증 연구(시기자 외, 2013a)의 후속 연구로 한국교육과정평가원에서 개발한 단문 형식의 쓰기 답안 채점에 적용 가능한 자동채점 프로그램의 성능에 대한 검증을 통해 공식적인 대규모 쓰기 단문형 평가에서 자동채점이 인간채점자 1명을 대체할 수 있는지 그 가능성을 탐색하고자 하였다. 이를 위해 쓰기 단문형 답안에 대해 인간채점자 2명에 의한 채점 자료와 자동채점이 인간채점자 1명을 대체하였을 때의 채점자 간 신뢰도, 채점자 엄격성, 검사점수의 일반화 가능도 계수, 채점 시간 및 비용 등에 대한 분석을 통해 자동채점 프로그램의 성능을 검증하였다.

현재 국가수준 학업성취도 평가의 서답형 문항은 채점위원들이 온라인상에서 채점하는 방식을 취하고 있으나, 전수 평가이기 때문에 채점 시간 및 비용, 채점 관리에 대한 부담이 매우 크다. 이러한 현실적인 문제로 인해 국가수준 학업성취도 평가의 서답형 문항은 주로 단답형 문항으로 구성되어 있다. 단문형 자동채점 프로그램의 타당성이 검증된다면 간단한 단어, 구 수준의 서답형 문항을 출제하고 있는 대규모 영어 평가의 채점에 소요되는 시간적·경제적·행정적 부담을 획기적으로 줄일 수 있을 것으로 예상된다.

Ⅱ. 단문형 쓰기 자동채점 프로그램 개관

1. 국내·외 단문형 자동채점 프로그램 개발 현황

가. 국외: c-rater, Auto-marking

국외에서 개발된 단문형 자동채점 프로그램으로 가장 잘 알려진 것은 ETS의 c-rater이며, 영국의 Cambridge 대학교에서도 단문형 답안을 채점하기 위해 Auto-marking이라는 정보추출기반의 자동채점 시스템을 개발한 바 있다.

1) c-rater

ETS에서 개발한 c-rater는 학생 답안이 의미 측면에서 정답과 얼마나 유사한가를 평가하여 점수를 부여하는 자동채점 프로그램이다. 여러 문장으로 구성된 에세이를 채점하는 대부분의 자동채점 프로그램과 달리 c-rater는 다소 짧게 구성되는 답안을 처리하며, 답안이 정답과 동일한 의미를 표현하고 있는지, 그렇지 않은지를 판단할 수 있도록 설계되었다(Sukkarieh &

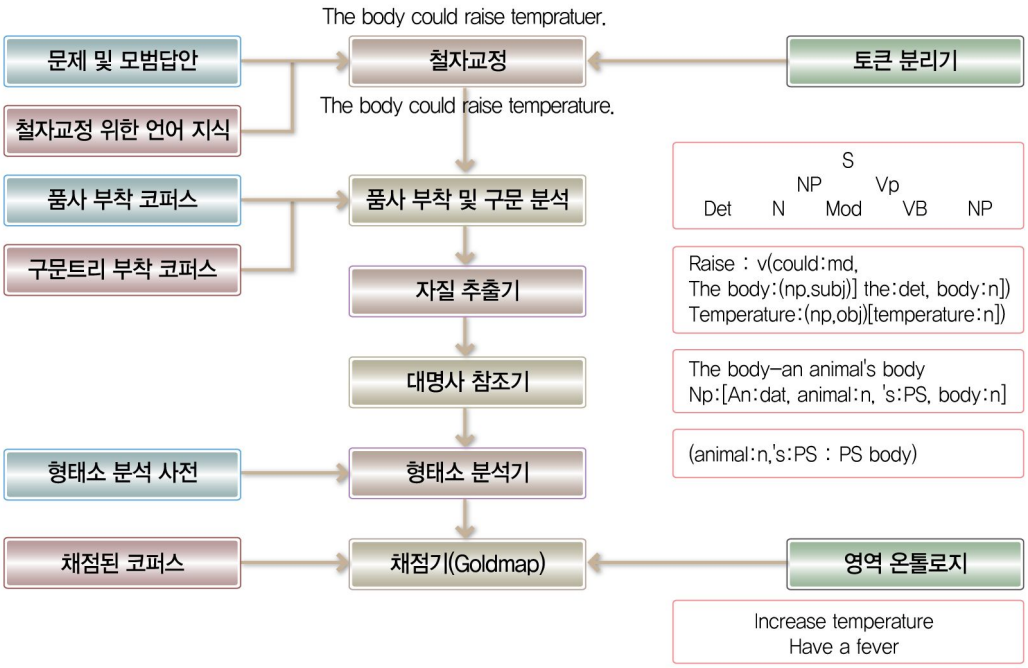
Blackmore, 2009). 그렇기 때문에 c-rater는 특정 교과목의 구체적인 내용의 이해 정도를 평가하도록 설계된 프로그램이다.

c-rater는 크게 모범 답안 구축, 언어 처리, 개념 인식(원문 함의), 점수 산출의 4단계 절차로 구성되어 있다. c-rater를 사용하여 자동채점을 실시하기 위해서는 우선 출제자가 문항의 정답모델을 구축해야 하며, 구축 과정을 용이하게 하기 위하여 'Alchemist'라는 인터페이스가 제공된다. 정답모델은 출제자가 제공한 정답 문장에서 개념만을 추출하여 정규화된 형태로 표현한 것이다. 이와 같은 정답 개념이 실제 문장으로 발현할 때에는 다양한 형태로 나타날 수 있다. 하나의 개념이 다양한 형태의 문장으로 발현되는 것을 바꾸어 말하기(paraphrase) 또는 원문 함의(textual entailment)라고 하며, c-rater는 바꾸어 말하기를 통해 표현된 다양한 변형 형태에 함의되어 있는 핵심 개념을 추출하여 정답모델과 비교하고 채점을 수행한다.

c-rater에서 핵심이 되는 채점자질은 바로 바꾸어 말하기에 함의되어 있는 핵심 개념을 수치화한 값이다. 초기 버전에서는 형태 일치 여부를 바탕으로 0, 1의 값을 채점자질로 사용하였다. 이와 같은 규칙 기반의 채점자질은 명료하고 계산 방식이 용이하지만 유연성이 떨어지는 단점이 있다. 따라서 최신 버전에서는 학생 답안의 표현과 모범 답안의 일치도를 확률값으로 표현하여 그 값이 0.5 이상이면 요구되는 개념을 가지고 있다고 판단하는 확률기반의 채점자질을 사용하고 있다(Sukkarieh, 2010; Sukkarieh & Blackmore, 2009).

c-rater는 자연어 처리 기법을 이용하여 학생 답안을 정답모델과 동일한 표준 형태(canonical representation)로 정규화한다. 이렇게 정규화된 표준 형태와 정답모델을 비교하고, 서로 일치된 정도에 따라 점수를 부여한다. 정규화 과정에서는 수동태/능동태와 같은 구문의 변형, 지시대명사의 참조 복원, 형태소 어형의 원형 복원, 동의어의 사용 등과 같은 변형을 정규화한다. 철자 오류도 정규화 대상에 포함되어 오류를 복원시킨 형태로 정규화된다. 이렇게 처리된 학생 답안은 자연어 처리 기법에 의한 구문 분석을 통해 동사 중심의 튜플(tuple)[2] 집합으로 표현되는데, 이로써 다양하게 나타나는 표층문이 정규화된 표현으로 변환된다. 다음으로 정규화된 학생 답안과 정답모델을 비교하는 규칙을 작성하고, 이를 이용하여 학생 답안이 정답과 얼마나 유사한지를 계산한다. 출제자의 정답모델 구축을 도와주는 Alchemist는 정답모델 구축 부분과 모델 검증 부분으로 구성되어 있다. Alchemist는 c-rater에 익숙하지 않은 출제자들이 정답에 포함되어야 할 채점 요소를 문장으로 입력하고 채점 요소와 유사한 단어, 동의어 등을 선택할 수 있도록 되어 있다. 이렇게 입력된 정답 문장은 자연어 처리 과정을 거쳐 튜플 집합의 구조를 지닌 정답모델의 구축에 사용된다. 모델 검증에서는 학생 답안을 c-rater로 자동채점하고, 인간 채점 점수와의 상관관계를 계산하여 검증하고 모델을 다시 수정한다(Leacock & Chodorow, 2003). c-rater의 시스템 구조는 〔그림 1〕과 같다.

2) 튜플은 여러 개의 항목을 연결시켜 제시하는 방식을 의미하며, "Walter wants money."라는 문장을 동사 중심의 튜플로 제시하면 "want: 주어 Walter: 목적어 money"가 된다.



〔그림 1〕 c-rater 시스템 구조

출처: Sukkarieh & Blackmore, 2009, p. 292.

c-rater의 성능은 미국의 NAEP(National Assessment of Educational Progress) 수학과 문제 논리 시험과 인디애나주의 11학년 영어 독해 시험에서 평가된 바 있다(Leacock & Chodorow, 2003). 두 시험에서 c-rater와 인간채점 간의 채점 일치도는 0.84로 나타나 c-rater의 성능이 상당히 우수하다는 사실이 확인되었다. 하지만 c-rater에도 점수 '미부여(miss)'와 점수 '과다부여(false positive)'의 문제는 존재한다. 미부여의 경우에는 자동채점하는 문항이 개방적일 때 그 비율이 증가하는 경향을 보였다. 한편 과다부여는 답안의 첫 번째 문장에 올바른 개념을 적시하였으나 이후의 문장이 주제와 다르게 작성된 경우와 답안에 사용된 단어들은 정확하지만 전체적으로 의미가 통하지 않게 작성된 경우에 발생하는 것으로 나타났다.

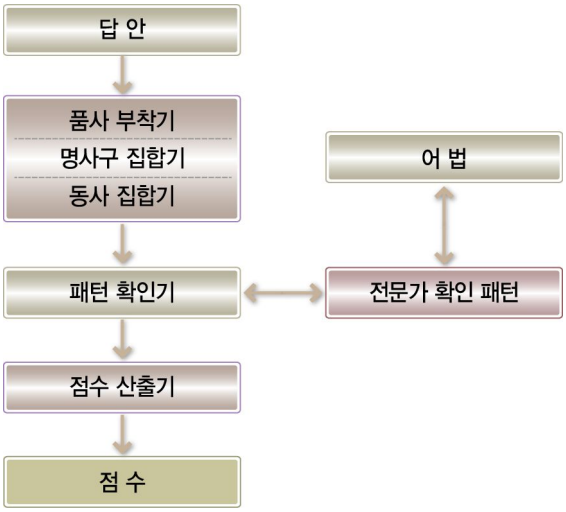
2) Auto-marking

Auto-marking은 Cambridge 대학교의 지역시험조합(University of Cambridge Local Examinations Syndicate)에서 시행하는 단문형 시험 문항을 자동으로 채점하기 위해 개발된 정보추출(information extraction) 기반 시스템이다(Sukkarieh, Pulman, & Raikes, 2003). 정보추출 기법은 완벽하고 정확한 구문분석을 필요로 하지 않으며 답안들이 비어법적이고 불완

전한 문장으로 구성된 경우에도 상당히 좋은 성능을 보이며 적용이 용이하다.

정보추출 기법에는 지식 공학(knowledge engineering)과 기계학습의 두 가지 접근 방식이 존재하는데, 지식 공학적 접근에서는 전문가가 정보의 패턴을 수동적으로 추출하는 반면에 기계학습적 접근에서는 시스템 자체가 정보의 패턴을 자동적으로 추출한다. 훈련용 답안이 많지 않을 경우에 주로 사용되는 지식 공학적 접근은 기계학습적 접근에 비해 정확성이 더 우수하다는 장점이 있지만 전문적인 지식과 기술을 갖춘 전문가의 시간과 노력이 전제되어야 한다. 이에 비해 훈련용 답안이 충분하며 채점 결과의 신뢰도가 크게 문제되지 않는 상황에서 사용할 수 있는 기계학습적 접근은 지식 공학적 접근에 비해 정확성이 다소 떨어지지만 인간전문가가 필요하지 않다는 장점이 있다.

정보 공학에 의한 정보추출 자동채점 시스템을 제시하면 〔그림 2〕와 같다(Sukkarieh & Pulman, 2005). 이 시스템의 첫 번째 단계에서는 품사 부착기와 명사구 및 동사류 집합기를 이용해서 답안의 패턴을 분석하며, 두 번째에서는 전문가에 의해 확인되었으며 어법상으로도 문제가 없는 패턴과 답안의 패턴을 비교하여 일치하는 패턴을 찾아낸다. 마지막으로 전문가의 패턴과 일치하는 답안의 패턴을 점수 산출기에 입력하게 되면 자동채점 점수가 산출된다.



〔그림 2〕 Auto-marking 시스템 구조

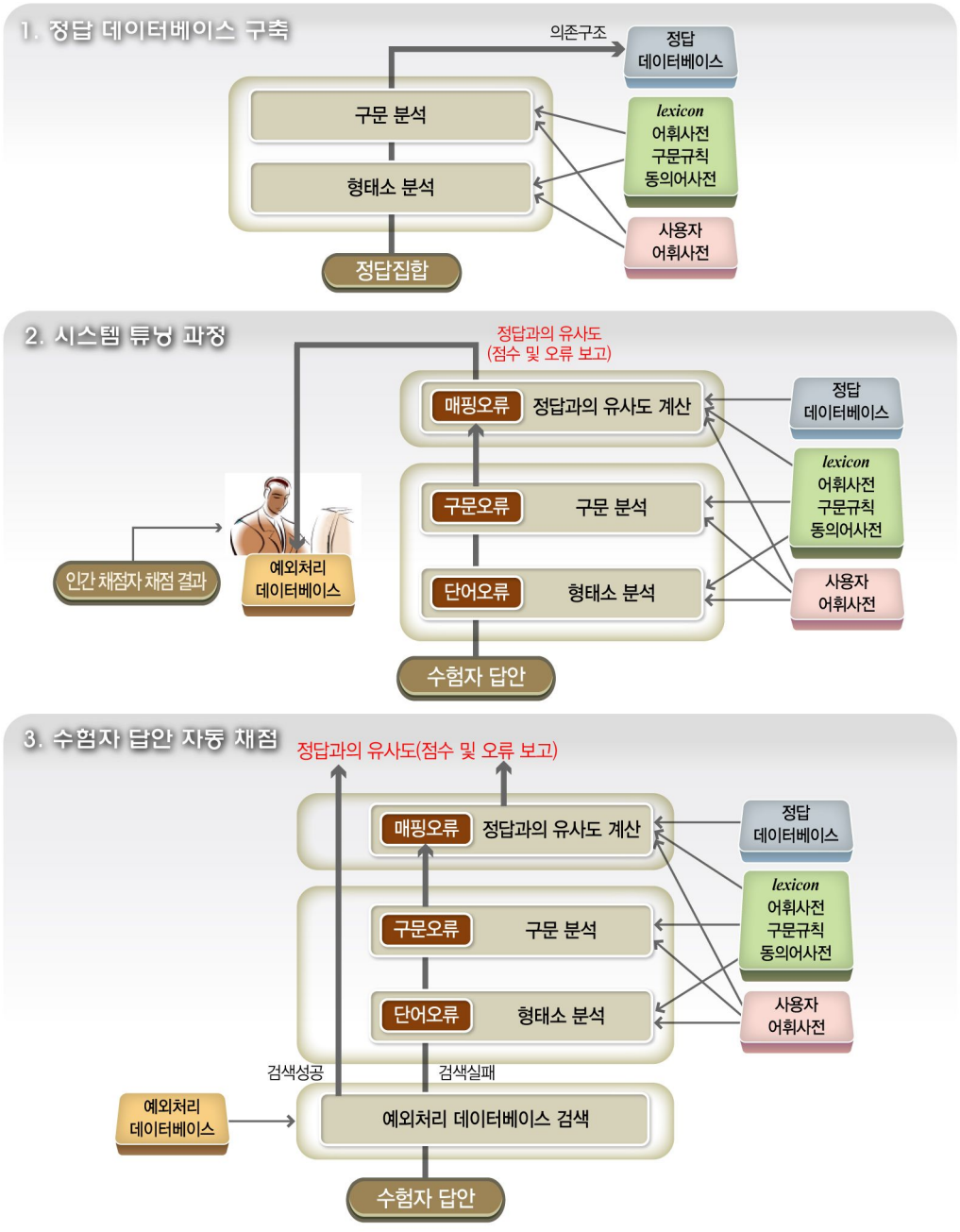
한편, 기계학습에 의한 정보추출 자동채점은 정보 공학적 접근에 비해 채점의 정확도가 낮은 편이다. 전반적으로 볼 때 문항이 요구하는 추론의 수준이 높을수록 자동채점의 성능은 떨어지며, 답안에 이중부정 문장이 사용된 경우나 모순되거나 불일치하는 정보가 포함된 경우에도 자동채점의 정확도가 낮아지는 것으로 밝혀졌다(Siddiqi & Harrison, 2008).

나. 국내: 진경애 외(2008), 노은희 외(2012)

현재까지 국내에서 개발된 단문형 자동채점 프로그램들 중에서 대표적인 것은 진경애 외(2008)의 영어 자동채점 프로그램과 노은희 외(2012)의 한국어 자동채점 프로그램이다.

1) 문장 단위 영어 자동채점 프로그램

진경애 외(2006, 2007, 2008)의 문장 단위 영어 자동채점 프로그램은 우리나라 학생들을 대상으로 통제형 영작문의 기본 문형을 익히고 피드백을 받을 수 있는 교수·학습용으로 활용할 수 있도록 개발되었다. 이 프로그램의 시스템은 크게 세 단계로 이루어져 있다. 첫 번째는 정답 데이터베이스를 구축하는 단계로 출제자가 제공한 정답 집합을 분석하여 정답 데이터베이스를 구축한다. 이와 같이 구축된 정답 데이터베이스는 학생 답안을 자동채점하는 데 사용된다. 두 번째 단계는 전체 시스템을 테스팅하고 튜닝하는 과정이다. 학생 답안에 대한 시스템의 처리 결과와 인간채점자의 채점 결과를 비교하여 시스템의 성능을 평가/튜닝하고 동시에 시스템의 처리 범위를 벗어나는 예외 문장에 대한 예외처리 데이터베이스를 구축한다. 이 단계는 인간채점자에 의해 수동으로 진행된다. 세 번째 단계는 앞의 두 단계에서 구축된 정답 데이터베이스와 예외처리 데이터베이스를 이용하여 학생 답안에 대한 자동채점을 수행하는 단계이다. 우선 학생 답안과 동일한 답안이 예외처리 데이터베이스에 존재하는 경우 데이터베이스의 채점 결과를 최종 결과로 보고한다. 그러나 검색이 실패하는 경우, 즉 예외처리 데이터베이스에 존재하지 않는 경우에는 학생 답안에 대해 형태소 분석과 구문 분석을 실시하여 학생 답안에 포함되어 있는 단어 오류와 구문 오류를 복원한다. 구문 분석이 수행된 학생 답안은 의존구조(dependency structure)로 표현되며, 정답 데이터베이스의 의존구조 중 가장 유사한 구조와의 비교를 통해 유사도를 계산한다. 마지막으로 구문 분석과 정답과의 유사도를 통해 최종 점수와 오류를 보고한다. 이 시스템의 구조를 나타내면 〔그림 3〕과 같다.



〔그림 3〕 문장 단위 영어 자동채점 시스템 구조

출처: 진경애 외, 2008, p. 43.

2) 한국어 단어・구 수준 서답형 문항 자동채점 프로그램

노은희 외(2012)의 한국어 단어・구 수준 서답형 문항 자동채점 프로그램(Korean Automatic Scoring System; KASS 1.0)은 정답이 3단어 이하인 한국어 서답형 문항을 자동채점하기 위해 개발되었다. 이 프로그램은 문항과 관련하여 미리 제시된 채점기준의 요구 사항을 컴퓨터로 인식하기 위하여 정답과 유사한 답안이나 오답 등의 정보를 자동채점에 적합한 형식으로 기술하고 이를 기반으로 자동채점을 실시하는 프로그램이다. KASS 1.0은 크게 자동채점 전처리 단계, 자동채점 단계, 자동채점 후처리 단계로 구분할 수 있다. 자동채점 전처리 단계는 자동채점을 하기 전에 학생 답안을 분석하고 정규화하는 단계이고, 자동채점 단계는 채점기준에 의거하여 작성된 정답 템플릿을 통해 일부의 학생 답안에 대하여 채점을 수행하고 정답과 오답으로 판정하기 어려운 답안에 대해서 미채점 답안으로 보류하는 단계이다. 자동채점 후처리 단계는 미채점 답안들의 클러스터링과 수작업 채점을 통해 미채점된 답안의 채점을 마무리하고 이에 따라 새로운 답안들의 자동채점 비율을 높이기 위해 채점기준을 수정・보완하고 수작업 채점 데이터베이스를 업데이트하는 단계이다. 이 프로그램은 일부 학생 답안의 채점 결과 분석 및 후처리 단계를 거쳐 정답 템플릿을 보완하여 더 많은 학생 답안에 대해서 자동채점을 반복적으로 가능하게 만드는 사이클형 구조로 설계되어 있기 때문에 대규모 평가의 자동채점 프로그램으로 적합성이 높다. KASS 1.0의 구조를 제시하면 〔그림 4〕와 같다.

다. 시사점

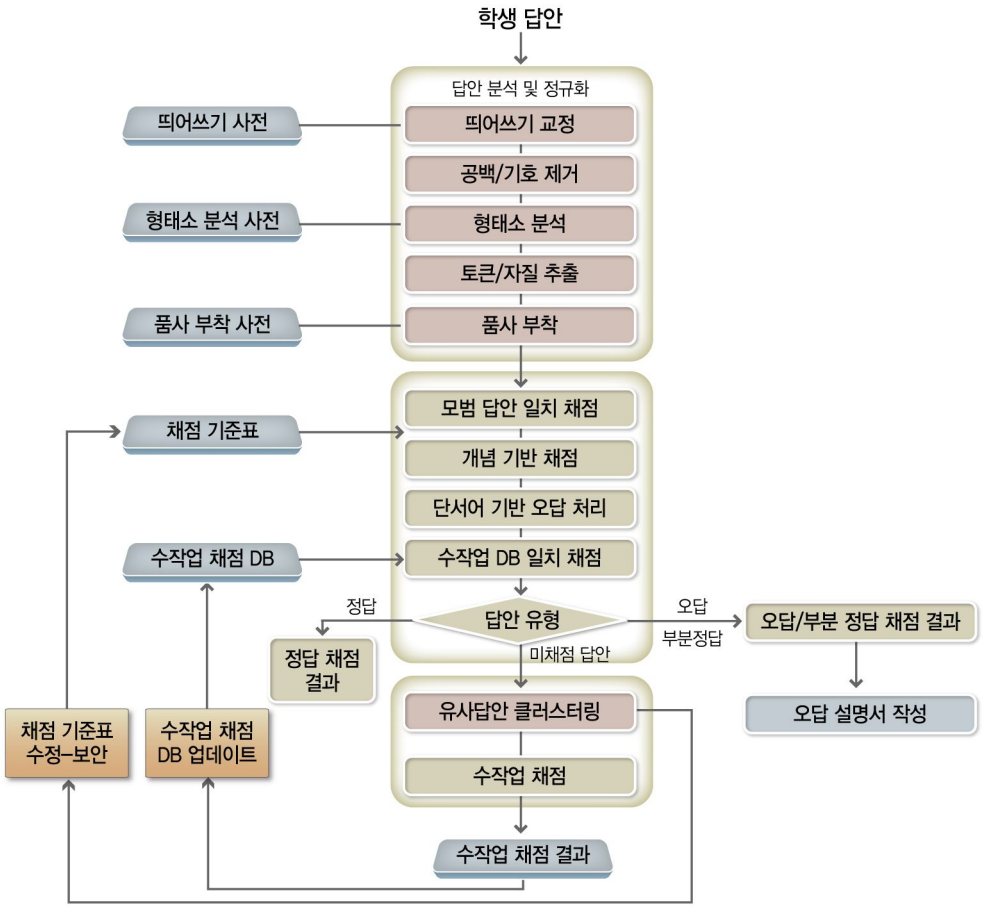
국내・외 자동채점 프로그램으로부터 얻을 수 있는 시사점을 정리하면 다음과 같다.

첫째, 답안에 대한 어휘・구문 분석과 오류 분석 등 전처리 과정에서는 에세이형 자동채점과 마찬가지로 분석 결과의 정확성이 매우 중요하다. 따라서 이에 요구되는 기능의 성능을 최대한 높일 수 있도록 프로그램을 개발해야 한다.

둘째, 단문형 자동채점에서는 개념 답안, 즉 정답모델을 구축하는 것이 매우 중요하다. c-rater의 Alchemist나 KASS 1.0의 정답 템플릿이 바로 정답모델을 구축하기 위한 모듈이며, 본 연구의 단문형 자동채점 프로그램에서도 역시 정답모델 구축을 도와주는 인터페이스를 개발해야 하고 더 나아가 정답모델 구축의 자동화를 적극적으로 반영할 필요가 있다.

셋째, 단문형 자동채점에 적합한 채점자질의 개발이 요구된다. 적용 원리가 분명한 규칙기반의 채점자질이 효율적인지 아니면 적용 범위가 폭넓은 통계기반의 채점자질이 효율적인지에 대한 교차 검토를 통해 단문형 문항 자동채점에 적합한 채점자질을 찾아내야 한다.

마지막으로 예외처리 데이터베이스화나 미채점 답안 클러스터링과 같은 방식을 통해 자동채점의 비율을 최대한 높일 필요가 있으며, 자동채점이 불가능한 답안을 시스템적으로 선별하는 방안도 적극적으로 고려해야 한다.

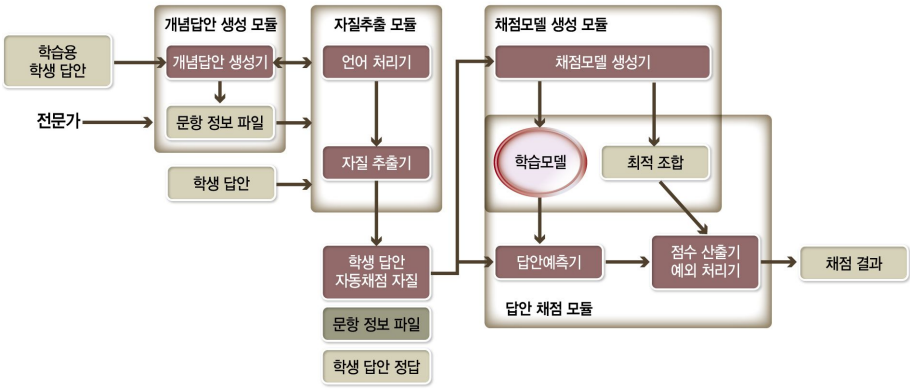


〔그림 4〕 한국어 단어·구 수준 서답형 문항 자동채점 시스템 구조

출처: 노은희 외, 2012, p. 82.

2. KICE 단문형 쓰기 자동채점 시스템 구성

한국교육과정평가원에서 개발한 단문형 자동채점 시스템은 '개념 답안 생성 모듈', '자질 추출 모듈', '채점모델 생성 모듈', '답안 채점 모듈'로 구성되어 있다. 각 모듈은 독립적으로 수행되며 여러 하위 프로그램들의 집합으로 구성된다. 단문형 자동채점 프로그램의 전체 구성은 〔그림 5〕와 같다.



〔그림 5〕 KICE 단문형 자동채점 프로그램 구성

가. 개념 답안 생성 모듈

채점 대상이 되는 문항 정보와 개념 답안을 생성하고 이를 저장한 '문항 정보파일'을 생성한다. 개념 답안 생성 모듈에서는 다음과 같은 작업을 수행한다.

- 고득점 학생들의 답안에서 의미를 가지는 단어(개념 답안 후보단어)를 추출하고, 유사한 답안들의 군집 생성
- 각 군집을 대표하는 개념 답안 추출
- 동의어 확장 프로그램과 전문가에 의해서 개념 답안 확장
- 확장된 개념 답안과 전문가에 의해 입력된 문항 정보를 저장한 문항 정보파일 생성

나. 자질 추출 모듈

자질 추출 단계는 입력된 학생들의 답안에 대한 언어처리와 자질 추출을 수행하여 자동채점 자질을 출력으로 제공한다. 자동채점자질은 채점모델 생성 모듈과 답안 채점 모듈에서 이용되며 총 37개의 자질이 추출된다. 단문형인 3급 1번과 3급 2번 문항은 2~4 문장으로 구성된 짤막한 답안을 요구하는 문항이다. 따라서 처리해야 하는 문장 및 요소가 적기 때문에 자동채점을 위해 추출할 수 있는 채점자질 또한 제한적이다. 단문형 자동채점을 위해 본 연구에서는 다음과 같이 채점자질을 선정하였다. (1) 채점의 효율성을 높이기 위해 채점기준 자체를 수치화하여 채점자질로 선정하였다. (2) 답안에 나타나야 할 핵심 단어들의 목록을 개념 답안으로 간주하여 자동으로 생성하고 이를 채점자질로 선정하였다. (3) 비속어와 노이즈 단어의 사용을 채점자질로 선정하였다. (4) 다양한 철자/구문 오류를 검출하여 이를 채점자질로 선정하였다. 답안의 길이가 짧은 단문형의 답안의 경우 철자 오류와 구문 오류를 정확하게 추출해 내는 것이 언어사용 영역의 성능과 직결된다. 본 연구에서는 우리나라 학생들의 영작문 능력을 고려하여 비문법적

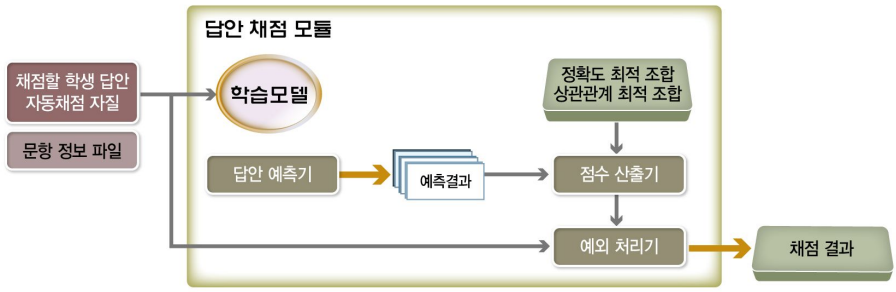
단어 나열에 대한 오류를 검출할 수 있는 방법을 도입하였다. 본 연구에서 선정한 채점자질은 총 37개로, 이 중 20개는 2012년에 개발한 에세이형 자동채점 프로그램에서 사용된 채점자질 중에서 선정하였고, 17개는 단문형 자동채점을 위해 새롭게 추가하였다. 본 연구에서는 이상의 채점자질을 모든 채점 영역에 공통으로 적용하였다. 자질 추출 모듈은 개념 답안 생성 모듈에서도 호출되는데, 이때는 고득점 학생들의 답안을 분석하여 기능어나 불용어를 제외한 단어의 원형을 출력으로 제공해 준다.

다. 채점모델 생성 모듈

채점모델 생성 모듈은 자질 추출 모듈에서 추출한 자동채점자질을 기계학습 알고리듬에 적용하여 채점모델을 생성한다. 채점모델 생성에 사용되는 기계학습 알고리듬에는 Sequential minimal optimization, Multilayer perceptron, Bayes network 알고리듬이 사용된다. 학습에 사용하는 점수는 전문 채점위원단이 부여한 기준점수를 이용한다. 채점모델 생성 모듈에서는 채점모델을 생성한 이후, 각 채점 영역별로 3가지의 알고리듬을 조합하여 가장 좋은 채점 결과를 나타내는 최적의 조합을 찾아낸다.

라. 답안 채점 모듈

답안 채점 모듈에서는 우선 자질 추출 모듈에서 추출한 자동채점자질을 적용하여 생성된 채점모델을 학습 모델에 입력으로 제공한다. 그리고 여러 개의 학습 모델의 결과를 최적으로 조합하여 최적의 점수를 생성한다. 본 연구에서는 Exact Matching(EM)과 Pearson Correlation(PC) 방법에 의해 생성된 자동채점 점수의 결과만을 제시하였다. 마지막으로 예외 처리 작업을 통해 최종 채점결과를 생성한다. 답안 채점 모듈의 구성은 〔그림 6〕과 같다.



〔그림 6〕 답안 채점 모듈 구성

Ⅲ. 연구 방법

1. 분석 자료

쓰기 자동채점 프로그램의 타당성을 검증하기 위해 사용된 자료는 2013년 국가영어능력평가시험 4월 예비평가와 6월 1차 평가 중 쓰기 3급 단문형 평가 자료이다. 분석에 사용한 단문형 문항의 특성과 예시 문항을 제시하면 다음과 같다.

〈표 1〉 단문형 문항의 특성

문항 번호	문항 유형	답안 제한 단어 수	답안 작성 시간
1	상황에 맞는 짧은 글쓰기	15~25 단어	5분
2	그림의 세부묘사 완성하기	하위문장별 10단어 이내	5분

예시 문항 상황에 맞는 짧은 글쓰기

시내에서 길을 잃은 외국인 친구에게 도움을 주어야 하는 상황이다. 다음에 제시된 세 가지 상황 중 하나를 선택한 후, 주어진 단어나 어구를 활용하여 상황에 알맞게 도움을 주는 메시지를 완전한 2~3개의 문장으로 작성하시오. (15~25 단어)
(1) information center/get
(2) a person/ask
(3) taxi/show

출처: 국가영어능력평가시험 2012년 3월 예비평가 3급 A형 1번 문항

예시 문항 그림의 세부 묘사 완성하기

아래의 글은 미용실에 있는 사람들의 행동을 묘사한 것이다. 다음 그림을 보고 (1)-(4)의 빈칸에 각각 10개 이내의 단어를 사용하여 문장을 완성하시오.
There are some people in a beauty salon.
The woman in the uniform is (1) _____.
The man in the uniform is (2) _____.
The lady in a scarf is (3) _____.
The boy on the sofa is (4) _____.

출처: 국가영어능력평가시험 2012년 5월 모의평가 3급 2번 문항

이 평가에서 학생들의 글쓰기 능력은 과제완성, 내용, 구성, 언어 사용의 네 가지 영역에서 평가되며, 채점 영역별로 점수를 산출하여 합산하는 분석적 채점 방식을 채택하고 있다. 2번

'그림의 세부 묘사 완성하기' 문항에서는 구성은 평가하지 않는다. 채점 영역별 주요 평가 내용은 〈표 2〉에 제시되어 있다. 무응답이거나 영어로 응답하지 않은 경우, 욕설을 쓰거나 문항이 제시한 주제와 상관이 없는 내용으로 답안을 작성한 경우에는 채점 불가 판정을 받으며 자동으로 0점이 부여된다.

〈표 2〉 단문형 문항의 채점 영역 및 평가 내용

채점 영역	주요 평가 내용
과제 완성 (Task Completion)	• 주어진 과제의 수행 정도 • 하위 문항의 수행 정도 • 주어진 주제에 맞는 내용의 적절성
내용 (Content)	• 내용의 충실성 • 내용의 구체성
구성 (Organization)	• 글의 논리적 조직성 • 글의 연결성/응집성
언어 사용 (Language Use)	• 다양한 문장 구조 및 어휘 사용 • 표현의 정확성 • 철자와 구두점의 정확성

평가 자료 중 일정 분량의 답안 및 채점 영역별 기준점수는 기계학습용으로 사용하였으며, 나머지 분량의 답안 및 이에 대한 채점 영역별 점수는 타당성 검증용으로 사용하였다. 기준점수는 전문 채점위원단에 의해 부여된 점수로 채점 점수의 정확성 판단을 위한 기준자료로 활용된다. 기준점수를 산출하는 목적은 자동채점의 기계학습용 데이터를 확보하는 것과 본채점에서 채점위원들의 신뢰도를 점검하는 것이다. 이를 위해 채점위원 인력풀에서 신뢰도가 높은 채점위원단을 확보하여 채점기준에 대한 워크숍을 실시하고, 기계학습에 필요한 일정 분량의 답안을 추출한 후 각 답안 당 세 명의 채점위원이 채점한 점수들의 최빈값을 산출하여 기준점수로 활용하였다. 본 연구에서 사용한 평가 자료는 〈표 3〉과 같다.

〈표 3〉 분석 자료의 구성

시험	문항	기계학습용	타당성 검증용	계
2013년 4월 예비	3급 1번	300	300	600
	3급 2번	297	297	594
2013년 6월 1차	3급 1번	301	301	602
	3급 2번	301	301	602

2. 분석 방법

쓰기 자동채점 프로그램의 타당성을 검증하기 위해 기준점수, 인간채점 및 자동채점에 따른 채점자 간 신뢰도, 검사점수 신뢰도, 채점자 엄격성을 비교하였으며, 자동채점의 효율성을 살펴보기 위해서 인간채점과 자동채점에 소요되는 채점 시간과 비용을 비교하였다. 자동채점 점수의 경우 Exact Matching(EM)과 Pearson Correlation(PC) 방법에 의해 산출된 점수를 비교하였다.

채점자 간 신뢰도는 인간채점 1과 2, 기준점수와 인간채점 평균, 기준점수와 자동채점 사이의 Pearson 적률상관계수, 유사일치도를 비교하였다. 채점자 간 신뢰도 비교에서 인간채점 평균은 2명의 평균 점수이기 때문에 0.5점 단위의 소수로 산출되지만, 정수로 처리한 결과와 거의 차이가 없어 분석 결과의 일관성과 간명성을 고려하여 반올림한 정수에 의한 결과만을 제시하였다. Pearson 적률상관계수는 -1~1의 범위를 가지며 1에 가까울수록 신뢰도가 높다고 볼 수 있다. 또한 일치도는 두 방법에 의한 채점 점수 중에서 그 값이 일치하는 점수들의 비율을 의미하며, 유사일치도는 일치도를 완화시킨 값으로서 두 방법에 의한 채점 점수들 중에서 인접한 점수를 가지는 점수들의 비율이다. 본 연구에서는 두 점수 간 ±1점 차이를 인접 점수로 간주하였다. 따라서 유사일치도가 1에 가까울수록 두 방법에 의한 채점 점수가 동일하다는 의미이며 신뢰도가 높다고 볼 수 있다.

다음으로 인간채점을 자동채점으로 대체하게 될 때 검사점수의 신뢰도가 어떻게 변화하는지 분석하기 위해 일반화가능도 이론(Brennan, 2001)을 적용하였다. 일반화가능도 분석에 적용한 설계는 채점 영역을 고정시킨 $p(\text{학생}) \times r(\text{채점자}) \times d(\text{채점영역})$ 교차 설계이다. 채점영역을 고정시킨 주된 이유는 국가영어능력평가시험의 평가틀에서 쓰기에 대한 채점 영역이 확정되어 있기 때문이며, 채점 영역에 따라 채점의 기준이 다르게 설정되어 있어서 특정 채점 영역의 점수를 다른 채점 영역의 점수로 일반화하는 데 한계가 있기 때문이기도 하다. 엄밀한 의미에서 볼 때 본 연구의 인간채점 자료는 여러 명의 인간채점자의 점수가 사용되었기 때문에 교차 설계에 해당되지 않는다. 그렇지만 채점자에 대한 훈련이 체계적이고 엄격하게 진행되었고 채점자 간 상관이 0.80을 훨씬 상회하였으므로 인간채점자들의 점수를 동일인의 점수로 가정하여 교차 설계로 분석하였다. 분석 결과에서 분산은 값이 클수록 검사점수에 미치는 영향이 크다는 의미이며, 일반화가능도 계수는 값이 1.0에 가까울수록 해당 채점 방식(인간채점 2명 또는 인간채점 1명+자동채점)에 의한 검사점수의 신뢰도가 높다고 볼 수 있다.

또한 기준점수, 인간채점, 자동채점의 엄격성을 비교하기 위해 다국면 라쉬 모형에 의한 로짓 척도상의 엄격성 지수를 ConQuest(Wu, Adams, Wilson, & Haldane, 2007) 프로그램으로 추정하여 비교하였다. 채점에서 엄격성 혹은 관대성은 채점자 및 채점 방법에 따라 다를 수 있으며 그 차이가 심할 경우 채점 결과의 신뢰도와 타당도를 저하시킬 수 있다. 논리적으로 볼 때 인간채점자에 비해 자동채점은 동일한 수준의 엄격성을 가지고 채점하기 때문에 채점의 신뢰

도를 제고하는 데 더 유리하지만, 인간채점과 자동채점의 엄격성이 큰 차이를 보이게 되면 채점의 타당도에 문제가 발생할 가능성도 존재한다(시기자 외, 2012). 엄격성 지수는 그 값이 양수이면 점수를 엄격하게 준다는 의미이며 음수이면 점수를 관대하게 준다고 해석할 수 있다. 마지막으로 자동채점의 효율성을 살펴보기 위해서 인간채점과 자동채점에 소요되는 채점 시간을 분석하였다.

Ⅳ. 연구 결과

1. 채점자 간 신뢰도

인간채점과 자동채점에 따른 채점자 간 신뢰도 비교에서는 인간채점 1과 2, 기준점수와 인간채점 평균, 기준점수와 자동채점 간의 Pearson 적률상관계수, 유사일치도를 비교하였다.

가. 상관계수

인간채점과 자동채점에 따른 상관분석 결과는 〈표 4〉와 같다.

4월 예비평가 단문형 문항에 대한 상관분석 결과, 기준점수와 인간채점자 평균 간 상관은 채점 영역에 따라 0.68~0.91로 산출되었고, 인간채점자 간 상관은 0.72~0.94로 분석되었다. 채점 영역에 따른 기준점수와 EM 방법, 그리고 기준점수와 PC 방법에 의한 자동채점 간 상관은 0.63~0.89로 나타났다. 6월 1차 평가의 경우, 기준점수와 인간채점자 평균 간 상관은 채점 영역에 따라 0.71~0.93으로 산출되었고, 인간채점자 간 상관은 0.62~0.80으로 분석되었다. 또한 채점 영역에 따른 기준점수와 EM 방법에 의한 자동채점 간 상관은 0.70~0.82, 기준점수와 PC 방법에 의한 자동채점 상관은 0.70~0.84로 나타났다.

문항별로 상관분석 결과를 살펴보면, 1번 문항의 경우 4월 예비평가에서는 과제완성 영역에서 기준점수와 인간채점자 평균 간 상관이 0.87로 가장 높았고, 다음으로 기준점수와 두 자동채점(EM과 PC) 간 상관이 높았다. 내용과 구성, 언어사용 영역에서는 기준점수와 인간채점자 평균 간 및 인간채점자 간 상관이 두 자동채점(EM과 PC)를 통해서 산출한 상관보다 높게 나타났다. EM과 PC의 두 자동채점 방법을 비교하면 모든 채점 영역에서 기준점수와 자동채점 간 두 방법 간에 매우 유사하게 산출되었다. 6월 1차 평가에서는 모든 채점 영역에서 기준점수와 인간채점자 평균 간 상관이 가장 높았고, 다음으로 기준점수와 두 자동채점(EM과 PC) 간 상관이 높게 나타났다. EM과 PC의 두 자동채점 방법을 비교하면 과제완성과 내용, 구성 영역에서 기준점수와 자동채점 간 상관이 PC 방법의 결과가 EM 방법의 결과보다 조금 높았다.

〈표 4〉 인간채점과 자동채점에 따른 상관계수 비교

문항		채점 영역	인간채점1 & 인간채점2	기준점수 & 인간채점 평균	기준점수 & 자동채점(EM)	기준점수 & 자동채점(PC)
3급 1번	4월	과제완성	0.83	0.87	0.84	0.84
		내용	0.72	0.68	0.64	0.65
		구성	0.78	0.73	0.63	0.63
		언어사용	0.82	0.87	0.72	0.73
		총점	0.89	0.91	0.81	0.81
	6월	과제완성	0.77	0.93	0.82	0.84
		내용	0.68	0.75	0.71	0.72
		구성	0.63	0.71	0.70	0.71
		언어사용	0.70	0.78	0.72	0.70
		총점	0.85	0.90	0.85	0.86
3급 2번	4월	과제완성	0.93	0.91	0.89	0.89
		내용	0.84	0.87	0.87	0.87
		구성	-	-	-	-
		언어사용	0.94	0.90	0.85	0.85
		총점	0.96	0.93	0.90	0.90
	6월	과제완성	0.79	0.86	0.75	0.75
		내용	0.62	0.75	0.71	0.71
		구성	-	-	-	-
		언어사용	0.80	0.81	0.81	0.80
		총점	0.87	0.88	0.82	0.82

2번 문항의 경우 4월 예비평가에서는 과제완성과 언어사용 영역에서 기준점수와 인간채점자 평균 간 및 인간채점자 간 상관이 두 자동채점(EM과 PC) 방법으로 산출한 상관보다 높았다. 내용 영역에서는 모든 채점자 간 상관이 0.84~0.87로 서로 유사하게 나타났다. EM과 PC의 두 방법 간 상관을 비교하면 모든 채점 영역에서 기준점수와 자동채점 간 상관이 EM과 PC 방법 모두 동일하였다. 6월 1차 평가의 경우 과제완성 영역에서는 기준점수와 인간채점자 평균 간 상관과 인간채점자 간 상관이 두 자동채점(EM과 PC)을 통해 산출한 상관보다 높게 분석되었다. 내용 영역에서는 기준점수와 인간채점자 평균 간 상관이 가장 높았고, 다음으로 기준점수와 두 자동채점(EM과 PC) 간 상관이 높게 나타났다. 언어사용 영역의 경우 기준점수와 인간채점자 평균 간 및 인간채점자 간 상관, 그리고 기준점수와 두 자동채점(EM과 PC) 간 상관이 0.80~0.81로 거의 차이를 보이지 않았다. EM과 PC의 두 방법 간 상관을 비교하면 과제완성

과 내용 영역에서는 두 방법에 의한 상관이 동일하였고, 언어사용 영역에서는 두 방법 간 상관이 매우 유사하였다.

나. 유사일치도

인간채점과 자동채점에 따른 유사일치도 분석 결과는 〈표 5〉와 같다.

〈표 5〉 인간채점과 자동채점에 따른 유사일치도 비교

문항		채점 영역	인간채점1 & 인간채점2	기준점수 & 인간채점 평균	기준점수 & 자동채점(EM)	기준점수 & 자동채점(PC)
3급 1번	4월	과제완성	0.95	0.98	0.97	0.97
		내용	0.63	0.50	0.64	0.64
		구성	0.72	0.62	0.63	0.63
		언어사용	0.97	1.00	0.93	0.93
	6월	과제완성	0.93	0.99	0.94	0.96
		내용	0.70	0.76	0.67	0.66
		구성	0.62	0.70	0.64	0.65
		언어사용	0.97	0.96	0.94	0.92
3급 2번	4월	과제완성	0.99	0.99	0.93	0.93
		내용	0.76	0.81	0.75	0.77
		구성	-	-	-	-
		언어사용	1.00	0.99	0.96	0.96
	6월	과제완성	0.99	0.99	0.93	0.93
		내용	0.72	0.82	0.75	0.75
		구성	-	-	-	-
		언어사용	1.00	1.00	0.94	0.92

4월 예비평가 단문형 문항의 유사일치도 분석 결과, 채점 영역에 따라 기준점수와 인간채점자 평균 간 유사일치도는 0.50~1.00, 인간채점자 간 유사일치도는 0.63~1.00으로 나타났다. 채점 영역에 따른 기준점수와 EM 방법에 의한 자동채점 간 유사일치도는 0.63~0.97, 기준점수와 PC 방법에 의한 자동채점 간 유사일치도는 0.63~0.97로 나타났다. 6월 1차 평가에서는 채점 영역에 따라 기준점수와 인간채점자 평균 간 유사일치도는 0.70~1.00, 인간채점자 간 유사일치도는 0.62~1.00으로 나타났다. 채점 영역에 따른 기준점수와 EM 방법에 의한 자동채점 간 유사일치도는 0.64~0.94, 기준점수와 PC 방법에 의한 자동채점 간 유사일치도는 0.65~0.96으로 나타났다.

문항별로 유사일치도 결과를 살펴보면 1번 문항의 경우 4월 예비평가에서는 과제완성 영역에서 모든 채점자 간 유사일치도가 0.95 이상으로 매우 높게 나타났다. 내용 영역에서는 기준점수와 두 자동채점(EM과 PC) 간 유사일치도가 가장 높았고, 구성과 언어사용 영역에서는 인간채점자 간 유사일치도가 가장 높게 산출되었다. EM과 PC의 두 자동채점 방법을 비교하면 두 방법 간 유사일치도는 모든 채점 영역에서 동일하게 나타났다. 6월 1차 평가에서는 과제완성 영역에서 모든 채점자 간 유사일치도가 0.93 이상으로 높게 나타났고, 그 중 기준점수와 인간채점자 평균 간 유사일치도가 0.99로 매우 높았다. 내용과 구성 영역에서도 기준점수와 인간채점자 평균 간 유사일치도가 다른 채점자 간 유사일치도에 비해 높았다. 언어사용 영역에서는 인간채점자 간 유사일치도가 0.97로 가장 높았고, 기준점수와 인간채점자 평균 간 유사일치도가 0.96, 기준점수와 EM 방법에 의한 자동채점 간 유사일치도가 0.94 순으로 높았다. EM과 PC의 두 자동채점 방법의 유사일치도를 비교해보면 모든 채점 영역에서 두 방법 간 유사일치도가 매우 유사하였다.

2번 문항의 경우 4월 예비평가에서는 과제완성 영역에서 모든 채점자 간 유사일치도가 0.93 이상으로 높게 나타났으며, 그 중 기준점수와 인간채점자 평균 간, 인간채점자 간 유사일치도가 모두 0.99로 매우 높았다. 내용 영역에서는 기준점수와 인간채점자 평균 간 유사일치도가 0.81로 가장 높았고, 다음으로 기준점수와 PC 방법에 의한 자동채점 간 일치도가 0.77로 높게 나타났다. 언어사용 영역의 경우 인간채점자 간 유사일치도가 1.00으로 매우 높았다. EM과 PC의 두 방법을 비교하면 과제완성과 언어사용 영역에서는 두 방법 간 유사일치도 차이가 없었으며, 내용 영역에서 PC 방법의 기준점수와 자동채점 간 유사일치도가 EM 방법의 유사일치도보다 높게 산출되었다. 6월 1차 평가의 경우 과제완성과 언어사용 영역에서 모든 채점자 간 유사일치도가 0.92 이상으로 높게 나타났다. 그 중 두 채점 영역 모두에서 기준점수와 인간채점자 평균 간 및 인간채점자 간 유사일치도가 두 자동채점(EM과 PC) 방법으로 산출한 유사일치도에 비해 높았다. 내용 영역에서는 기준점수와 인간채점자 평균 간 유사일치도가 가장 높았고, 인간채점자 간 유사일치도는 가장 낮았다. EM과 PC의 두 방법 간 유사일치도를 비교하면 과제완성과 내용 영역에서는 두 방법 간 유사일치도 차이가 없었으며, 언어사용 영역에서 EM 방법의 유사일치도가 PC 방법의 유사일치도에 비해 조금 높게 분석되었다.

2. 검사점수 신뢰도

채점 영역이 고정된 p(학생)$\times$ r(채점자)$\times$ d(채점영역) 설계의 일반화가능도 분석을 통해 인간채점자 2명에 의한 채점 상황 및 인간채점자 1명과 자동채점에 의한 채점 상황에 따른 학생 국면과 채점자 국면의 분산과 일반화가능도 계수를 비교하였다. 인간채점자 2명에 의한 채점 상황, 인간채점자 1명을 자동채점(EM과 PC)으로 대체한 채점 상황에 따른 국면별 분산

과 일반화가능도 계수는 〈표 6〉에 제시하였다.

검사점수의 신뢰도에 대한 분석 결과, 검사점수의 차이는 주로 학생 능력에 의해 발생하며, 인간채점자 1명을 자동채점으로 대체할 경우 검사점수의 신뢰도가 0.92 이상의 높은 수준으로 유지되는 것으로 나타났다.

〈표 6〉 인간채점과 자동채점에 따른 검사점수 신뢰도 비교

문항		국면	인간채점1 & 인간채점2	Exact Matching (EM)		Pearson Correlation (PC)	
				인간채점1 & 자동채점	인간채점2 & 자동채점	인간채점1 & 자동채점	인간채점2 & 자동채점
3급 1번	4월	p(학생)	0.75	0.63	0.67	0.64	0.68
		r(채점자)	0.03	0.07	0.06	0.07	0.06
		pr(학생×채점자)	0.03	0.03	0.03	0.03	0.03
		일반화가능도 계수	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**
	6월	p(학생)	0.53	0.44	0.39	0.46	0.41
		r(채점자)	0.05	0.07	0.06	0.07	0.07
		pr(학생×채점자)	0.03	0.02	0.02	0.02	0.02
		일반화가능도 계수	**0.95**	**0.95**	**0.94**	**0.95**	**0.95**
3급 2번	4월	p(학생)	1.30	1.23	1.26	1.23	1.27
		r(채점자)	0.00*	0.00	0.00	0.00	0.00
		pr(학생×채점자)	0.02	0.03	0.03	0.03	0.03
		일반화가능도 계수	**0.99**	**0.97**	**0.97**	**0.97**	**0.97**
	6월	p(학생)	0.50	0.51	0.46	0.52	0.47
		r(채점자)	0.00	0.00	0.00	0.01	0.01
		pr(학생×채점자)	0.02	0.03	0.04	0.03	0.04
		일반화가능도 계수	**0.95**	**0.95**	**0.92**	**0.95**	**0.92**

* 분산 중에서 0.00은 대부분 반올림 때문에 발생함.

3. 채점자 엄격성

단문형 문항에 대한 인간채점, 기준점수, 자동채점(EM과 PC)의 엄격성은 〈표 7〉과 같다.

4월 예비평가의 채점자 엄격성에 대한 분석 결과, 3급 1번에서는 -0.81~1.87, 3급 2번에서는 -0.89~0.80의 상당히 큰 변동 폭을 보였다. 한편 기준점수의 엄격성은 두 문항 모두에서 0.50 미만의 작은 값을 보였지만, 자동채점의 엄격성은 -0.89~-0.78로 나타나 다소 관대한 채점 경향을 보였다. 6월 1차 평가에서도 인간채점의 엄격성은 변동의 폭이 상당히 크게 나타

나, 3급 1번에서는 -0.81~1.59, 3급 2번에서는 -1.46~1.38의 범위를 보였다. 한편 기준점수의 엄격성은 두 문항 모두에서 0.50 미만의 작은 값을 보였고, 자동채점의 엄격성도 대부분 -1.0 미만으로 작았지만 PC 방법에 의한 자동채점은 엄격성이 -1.20으로 나타나 상당히 관대한 채점 경향을 보였다.

〈표 7〉 인간채점과 자동채점에 따른 채점자 엄격성 비교

시험	채점자		엄격성(logit)	
			3급 1번	3급 2번
4월 평가	인간채점	인원	6	5
		최솟값	-0.81	-0.89
		최댓값	1.87	0.80
	기준점수		0.37	0.20
	자동채점(Exact Matching)		-0.81	-0.78
	자동채점(Pearson Correlation)		-0.81	-0.89
6월 평가	인간채점	인원	11	11
		최솟값	-0.81	-1.46
		최댓값	1.59	1.38
	기준점수		-0.04	0.20
	자동채점(Exact Matching)		-0.50	-0.69
	자동채점(Pearson Correlation)		-0.12	-1.20

4. 효율성 평가

자동채점 프로그램을 국가영어능력평가시험에 도입한 주요 이유는 인간채점에 비해 채점에 소요되는 시간과 비용이 획기적으로 감소되기 때문이다. 여기에서는 실증적인 자료를 비교·분석하여 자동채점 프로그램의 효율성이 어느 정도인지를 파악하였다.

가. 채점 소요 시간

학습 데이터를 이용한 채점모델 생성 시간과 자동채점에 소요되는 시간을 측정한 결과는 〈표 8〉과 같다.

〈표 8〉 단문형 자동채점 프로그램의 채점 소요 시간

(단위: 초)

문항		채점모델 생성				자동채점		
		시험 및 답안 개수	언어처리 및 자질추출	기계학습	합계	언어처리 및 자질추출	기계학습	합계
3급 1번	4월	300개 답안당	10 0.03	3011.41 10.04	3021.41 10.07	10 0.03	4.16 0.01	14.16 0.04
	6월	301개 답안당	10 0.03	3008.70 10.00	3018.70 10.03	10 0.03	4.19 0.01	14.19 0.04
3급 2번	4월	297개 답안당	10 0.03	2852.87 9.61	2862.87 9.64	10 0.03	4.04 0.01	14.04 0.04
	6월	301개 답안당	10 0.03	2959.83 9.93	2969.83 9.96	10 0.03	4.15 0.01	14.15 0.04

채점모델 생성 시간은 학습 데이터를 이용하여 과제완성, 내용, 구성, 언어사용 채점 영역에 대한 3가지 알고리듬의 채점모델과 최적 조합을 생성하는 데 소요된 시간이고 자동채점 시간은 4가지 채점 영역에 대한 예측 점수와 예외처리를 통해 최종 정답을 도출하는 데 소요되는 시간이다.

3급 1번과 3급 2번과 같은 단문형 문항은 답안의 길이가 에세이형에 비해 짧기 때문에 채점에 소요되는 시간도 함께 짧아진다. 따라서 인간채점에 소요되는 단문형 문항의 답안당 채점 소요 시간을 에세이형 문항에 대한 소요 시간 151.6초의 2/3(101.07초)로 산정하여 인간채점과 자동채점에 따른 채점 소요 시간을 비교하면 〈표 9〉와 같다.

〈표 9〉 인간채점과 자동채점(단문형)에 따른 채점 소요 시간

구분		소요 시간(A)	답안 개수(B)	답안당 소요 시간
인간채점		-	-	101.07
자동채점 (3급 1번)	모델 생성	12159.75	1,209	10.06
	자동채점	56.59	1,209	0.046
자동채점 (3급 2번)	모델 생성	11964.79	1,206	9.92
	자동채점	56.06	1,206	0.046

나. 채점 비용[3)]

국가수준 학업성취도평가와 같은 대규모 평가 서답형 문항에 대한 답안 채점에서는 채점의 신뢰도를 확보하기 위해 2명의 채점자가 복수 채점을 하도록 설계되어 있다. 따라서 대단위 시

3) 본 절은 시기자 외(2012, pp. 154-155)의 내용을 발췌한 것임.

험이 시행될 경우 인간채점은 채점 수당에 따른 경제적인 부담을 피할 수 없다. 현재 개발 중인 쓰기 자동채점 프로그램을 실제 채점에 적용 시 기준점수 산출 비용이 직접적으로 발생하고, 간접비용으로 시스템 관리 및 유지 보수 비용이 발생할 것으로 예상된다. 이와 비교하여 온라인으로 진행되는 인간채점은 많은 직·간접 비용이 발생할 것으로 예상된다. 1개 답안 당 1,000원의 채점 수당을 받는다고 할 때, 약 60만명의 학생 답안에 대한 채점 수당을 계산하면 12억원(60만×1000원×2명) 정도가 소요된다. 뿐만 아니라 채점자 간의 점수가 일정 기준 이상 차이가 날 경우 대규모 평가에서는 대부분 재채점을 시행하고 있고 이에 따른 추가 비용이 발생하게 된다. 이와 같은 비용은 순수하게 채점 수당만을 계산한 것이며, 채점 관리 시스템 유지 보수 및 채점 지원 요원에 대한 수당과 같은 간접비용까지 포함될 경우 인간채점은 비용적인 측면에서 자동채점에 비해 매우 경제성이 낮은 것으로 판단된다.

Ⅴ. 결론 및 제언

본 연구에서는 대규모 영어 단문형 쓰기 평가에서 자동채점 프로그램의 적용 가능성에 대해 알아보기 위하여 인간채점과 자동채점에 따른 채점자 간 신뢰도, 채점자 엄격성, 검사점수 신뢰도, 채점 소요 시간 및 비용 등의 차이를 비교하였다. 분석 결과를 요약하면 다음과 같다.

첫째, 인간채점과 자동채점에 따른 채점자 간 신뢰도를 상관계수, 유사일치도에 의해 비교한 결과, 전반적으로 기준점수와 자동채점의 채점자 간 신뢰도는 기준점수와 인간채점 평균 및 인간채점자 2명의 채점자 간 신뢰도와 유사한 수준으로 나타났으며, 특정 문항의 과제완성 영역에서는 오히려 기준점수와 자동채점의 채점자 간 신뢰도가 가장 높은 경우도 발견되었다. 한편 EM과 PC 방법에 의한 자동채점의 채점자 간 신뢰도에서는 별다른 차이가 나타나지 않았다. 따라서 단문형 자동채점과 인간채점자 1명의 채점자 간 신뢰도는 인간채점자 2명의 채점자 간 신뢰도와 유사한 수준으로 유지될 수 있음을 알 수 있다.

둘째, 인간채점자 1명을 자동채점으로 대체하였을 때 검사점수 신뢰도의 변화를 살펴본 결과, 자동채점이 인간채점자 1명을 대체하는 경우 검사점수의 신뢰도는 인간채점자 2명에 의한 검사점수 신뢰도와 같이 0.90 이상으로 유지되었으며, EM과 PC 방법에 의한 자동채점 검사점수 신뢰도는 거의 유사하였다. 따라서 자동채점이 인간채점자 1명을 대체하더라도 검사점수의 신뢰도는 상당히 높게 유지될 수 있을 것으로 예상된다.

셋째, 인간채점과 자동채점에 따른 채점자의 엄격성을 비교한 결과, 인간채점의 엄격성은 큰 차이를 보이는 반면, 기준점수의 엄격성은 전반적으로 0에 가까운 값을 보여 채점 결과에 대한 엄격성의 영향력이 크지 않았다. 하지만 자동채점의 엄격성은 3급 2번에 대해서 관대한 채점

경향을 보였는데, 이것은 3급 2번과 같이 짧은 문장을 완성하는 문항의 경우 인간채점자들이 어법에 초점을 맞추어 엄격하게 채점하기 때문인 것으로 판단된다.

넷째, 인간채점과 자동채점에 의해 소요되는 채점 시간과 비용을 비교해본 결과, 단문형 문항의 답안당 채점 시간은 인간채점이 101.07초이지만 자동채점은 10초 내외로 1/10에 불과하였으며, 비용면에서도 인간채점보다 경제적인 것으로 확인되었다.

단문형 자동채점 프로그램의 성능을 향상시키기 위한 후속 연구를 제안하면 다음과 같다.

첫째, 개념 답안 추출 방법에 대한 개선이 필요하다. 개념 답안 추출 시 고득점을 받은 수험생의 답안으로부터 자동으로 추출하는 방법을 적용해 본 결과, 현재는 수험생의 답안에 가장 많이 출현한 내용어를 중심으로 개념 답안이 추출되는 경향을 보였다. 그러나 답안의 길이가 길어질 경우 출현 빈도만 고려하게 되면 정확도는 감소될 것으로 예상된다. 따라서 개념 답안 추출 방법을 더 세분화할 필요가 있다.

둘째, 문장의 복잡성에 근거한 구문 오류 및 철자 오류 검출 방식에 대한 고려가 필요하다. 본 연구에서 문장의 복잡성을 구문 오류 및 철자 오류 검출에 적용하여 본 결과, 기존의 오류 검출 방식보다 검출력이 우수한 것으로 나타났으며, 이를 더 확장하여 구현할 만한 가치가 있는 것으로 확인되었다.

셋째, 채점자질에 대한 선택이나 분류 방안에 대한 연구가 필요하다. 기계학습에 의한 채점점수와 각 채점자질 값과의 상관계수를 산출하여 채점자질별 중요도를 분석해 본 결과, 문항 유형에 따라 채점자질의 중요도가 다르게 나타나는 것으로 확인되었다. 이러한 결과는 인간채점자가 중요하다고 생각하는 채점자질이 실제 자동채점에서는 예측한 방향대로 기능하지 않을 수 있음을 보여 준다. 따라서 채점자질의 중요도를 판단할 때 인간채점자와 자동채점 간의 차이를 최소화할 수 있는 방안을 모색할 필요가 있다.

참 고 문 헌

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

시기자, 박도영, 이용상, 박상욱, 임은영, 구슬기, 임황규, 최연희, 이공주, 김지은, 김성, 이은숙, 김성묵, 윤경아, 이순웅(2012). 국가영어능력평가시험 쓰기 자동채점 프로그램 개발. 한국교육과정평가원 연구보고 RRE 2012-10.

시기자, 이용상, 박도영, 임황규, 구슬기, 박상욱, 임은영(2013a). 국내 대규모 영어 쓰기 평가에서의 자동채점의 적용 가능성 탐색. **교육평가연구, 26**(2), 319-345.

시기자, 박도영, 임황규, 최연희, 표경현, 이공주, 이은숙, 윤경아, 성열원, 이순웅(2013b). 국가영어능력평가시험 쓰기 자동채점 프로그램 개발(Ⅱ). 한국교육과정평가원 연구보고 RRE 2013-11.

진경애, 남명호, 김명화, 오상철, 김민정, 주형미, 신효필, 반재천, 김수경(2006). 서답형 문항 자동채점 프로그램 도입 방안 연구(Ⅰ). 한국교육과정평가원 연구보고 RRI 2006-6.

진경애, 이병천, 주형미, 신동광, 박정, 김지은, 이공주, 이은성(2007). 서답형 문항 자동채점 프로그램 도입 방안 연구(Ⅱ): 영작문 채점을 중심으로. 한국교육과정평가원 연구보고 RRE 2007-4.

진경애, 이병천, 신동광, 박태준, 주헌우(2008). 서답형 문항 자동채점 프로그램 도입 방안 연구(Ⅲ). 한국교육과정평가원 연구보고 RRE 2008-6.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Leacock, C., & Chodorow, M. (2003). c-rater: Automated scoring of short-answer questions. *Computers and humanities, 37*(4). 389-405.

Siddiqi, R., & Harrison, C. J. (2008). *On the automated assessment of short free-text responses*. Paper presented at the 34th International Association for Educational Assessment (IAEA) Annual Conference, Cambridge, UK.

Sukkarieh, J. Z. (2010, July). *Using a MaxEnt classifier for the automatic content scoring of free-text responses*. Paper presented at the 30th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Chamonix, France. Retrieved from http://www.lss.supelec.fr/MaxEnt2010/ paper /048.pdf.

Sukkarieh, J. Z., & Blackmore, J. (2009). *c-rater: Automatic content scoring for short constructed response*. In Proceedings of the 12th International Conference on Artificial

Intelligence in Education, Amsterdam, Netherlands.

Sukkarieh, J. Z., & Pulman, S. G. (2005). Automatic short answer marking. In *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, June 2005*. Association for Computational Linguistics.

Sukkarieh, J. Z., Pulman, S. G., & Raikes, N. (2003). *Auto-marking: Using computational linguistics to score short, free text responses*. Paper presented at the 29th annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Wu, M., Adams, R., Wilson, M., & Haldane, S. (2007). *Conquest 2.0*. Camberwell: ACER.

• *논문접수 : 2014-05-01/ 수정본접수 : 2014-06-01/ 게재승인 : 2014-06-13*

ABSTRACT

Applicabilities of Automated Short-answer Scoring to Large-scale English Writing Tests

Ki-Ja Si
(Research fellow, Korea Institute for Curriculum and Evaluation)
Do-Young Park
(Associate Research fellow, Korea Institute for Curriculum and Evaluation)
Hwang-Gyu Lim
(Researcher, Korea Institute for Curriculum and Evaluation)

This study seeks to the possibilities of applying automated short-answer scoring to large-scale English writing tests through verifying the performance of an automated short-answer scoring program customized for the Level 3 Writing Section of National English Ability Test (NEAT). Two items of the NEAT Level 3 Writing Section, that is 'Wring a short story' and 'Completing four detailed picture descriptions' require short answers which are scored analytically. To verify the performance of automated scoring, differences of automated and human scoring in correlations, agreement indices, rater severities, generalizability coefficients, scoring time and expenses were investigated. Results revealed that automated scoring could maintain inter-rater and test-score reliabilities as high as human scoring. Furthermore, automated scoring drastically reduced the scoring time and expenses as well as the fluctuation of rater severity.

Key Words : short-answer writing tests, inter-rater reliability, rater severity, generalizability coefficients