교육과정평가연구
The Journal of Curriculum and Evaluation
2014, Vol. 17, No. 2, pp. 99~122

한국어 서답형 문항 자동채점 결과 비교 분석[1)]
- 국가수준 학업성취도 평가 국어, 사회, 과학 문항을 중심으로 -

노 은 희(한국교육과정평가원)*
성 경 희(한국교육과정평가원)**

≪ 요 약 ≫

본 연구는 2013년 개발된 한국어 서답형 문항 자동채점 프로그램을 활용하여 2012년 학업성취도 평가의 초·중·고 국어, 초·중 사회/과학의 총 38문항 각 3,010개 답안을 대상으로 교과 간 문항 및 답안 유형의 차이, 교과 간 자동채점 결과의 차이를 분석하였다.

우선, 2009~2012년 학업성취도 평가 서답형 답안의 유형별 비율은 단어·구 답안이 74.7%로 가장 높았고, 문장 답안이 12%, 다문장 답안이 10.9%, 기타 답안이 2.4%를 차지하였다. 교과별로 살펴보면, 단어·구 답안(P1~P3)의 경우 사회 교과가 86.5%로 가장 높았으며, 문장 답안(P4~P6)은 국어 교과가 17.0%로 가장 높았다. 기타 답안(그래프, 선긋기 등)의 경우 과학 교과가 6.6%로 상대적으로 비율이 높았으며, 국어, 사회 교과에서는 거의 출제되지 않았다. 즉, 국어 교과에서는 술어형 단어나 구, 문장 형태의 답안을, 사회 교과에서는 내용함축적 개념어 형태의 답안을 요구하는 문항이 자주 출제되었다. 다음으로 자동채점 결과, 단어·구 수준 서답형 문항의 Kappa계수는 최소 .95 이상으로 채점 신뢰도가 매우 높게 나타났으나, 답안의 길이가 증가하고 복잡해질수록 인간채점과 자동채점 간 일치도가 떨어지는 것으로 나타났다. 채점 비율 측면에서는 국어 문항이 평균 99.73%로 가장 높았으나, 채점 신뢰도 측면에서는 사회 문항이 가장 높은 신뢰도(Kappa계수 평균 1.00)를 보여주었다.

요컨대 국어, 사회, 과학의 교과별 사용 용어 및 용례, 문항 출제 형식은 서로 다른 특징을 보이며, 이는 채점 결과에도 영향을 미쳤다. 이를 볼 때, 교과별로 지식베이스를 구축하고 이와 연계되어 차별화된 자연언어처리 및 개념 분석 기술이 정교화된다면, 현재의 단어·구 수준 자동채점 프로그램의 채점 정확성 및 효율성을 상당한 정도로 높일 수 있을 것으로 기대된다.

주제어: 자동채점, 한국어 자동채점 프로그램, 서답형 문항, 국가수준 학업성취도 평가

1) 본 연구는 한국교육과정평가원(2013)에서 수행한 '대규모 평가를 위한 서답형 문항 자동채점 프로그램 정교화 및 시범 적용'의 일부 내용을 재구성한 것임.

* 제1저자, noro@kice.re.kr
** 교신저자, kelly9147@kice.re.kr

Ⅰ. 서론

오늘날 교육 평가 분야에서 컴퓨터 및 인터넷을 활용하는 평가 체제의 도입과 서답형 문항 활용에 대한 강조는 세계적인 추세이다(성태제 외, 2013; 김경희 외, 2013). 이에 국내 교육계에서도 컴퓨터 기반 대규모 평가의 가능성을 탐색하는 한편, 서답형 문항의 비중 확대를 적극적으로 고려하고 있다(김경희 외, 2013; 교육과학기술부, 2010, 2011).

서답형 문항은 수험자의 자유로운 응답을 허용하기 때문에 창의력, 문제해결력, 비판력, 논리적 일관성 등의 고등정신 능력을 평가하는 데 효과적인 반면에, 채점 비용과 채점 결과의 공정성 문제로 대규모 평가에서 널리 사용하기 어려운 측면이 있다.[2] 또한 서답형 문항은 간단한 기입형에서부터 에세이 형태까지 다양한데, 에세이 같은 경우는 문법적 표현에 대한 간단한 점검부터 글의 구성과 주제 구현 능력까지를 판단해야 하기 때문에 전문가가 채점을 하는 데도 상당한 어려움이 따른다. 그러나 대규모 평가의 경우 모범 답안과의 일치 정도를 판단하여 채점하는 간단한 기입형 문항(단어・구 수준)이 다수를 차지하고 있어서 컴퓨터 자동채점의 가능성이 매우 큰 편이므로, 자동채점 개발 및 연구는 많은 인력과 시간이 투입되는 현재 채점 체제의 비용 과다 문제를 상당 정도 해결할 수 있다.

한국교육과정평가원에서는 대규모 평가에서 서답형 문항의 활용을 제고하고 채점의 효율화를 도모하기 위해, 2012, 2013년에 걸쳐 단어・구 수준의 한국어 서답형 문항 자동채점 프로그램을 개발・적용하는 연구를 진행하였다(노은희 외, 2012; 2013). 이에 본 연구는 다양한 교과에 대한 프로그램의 적용 가능성과 성능을 살펴보기 위해, 2013년 개발된 한국어 서답형 문항 자동채점 프로그램 KASS(Korean Automatic Scoring System)을 활용하여 학업성취도 평가에서 한국어로 작성되는 국어, 사회, 과학 교과의 문항을 검증하고자 한다.

그런데 대규모 평가에서 자동채점 프로그램의 외연적 확장을 꾀하고자 할 때, 다음과 같은 세 가지 이유로 교과별 특수 상황을 고려하지 않을 수 없다. 첫째, 해당 교과가 기반하고 있는 학문 분야에 따라, 또한 해당 학문적 내용이 교육적 상황에서 어떻게 변환되어 교수되는지 그 맥락에 따라 교과마다 주로 사용하는 용어 데이터풀이 다르다는 점이다. 예를 들어, 과학 교과의 '뿌리털', '이산화탄소(CO_2)', '대뇌', 그리고 사회 교과의 '피고인', '민사소송', '시장가격', '사헌부' 등의 용어들은 해당 교과에서 전용되는 용어로 기타 교과에서 서답형 답안으로 자주 활용하

2) 현재 국가수준 학업성취도 평가의 서답형 문항 채점 방식인 온라인 인간채점의 경우, 2012년에 고2 50만 명 응시자의 국어, 수학, 영어 답안 채점과 관련하여 7월 19일부터 8월 14일까지 한 달여 동안 채점자 3,689명, 관리자 115명이 동원되어 총 비용이 약 14억 원 가량이 소요되었다(김미경 외, 2012). 2011년 초6, 중3의 경우 1,210만 명 응시자의 국어, 수학, 영어, 사회, 과학 답안 채점과 관련하여 시도별로 평균 12.7일 동안 7,366명이 동원되어 총 비용이 약 49억 원 가량이 소요되었다(김경성 외, 2011). 즉, 한 해 초・중・고 학업성취도 평가의 서답형 채점으로 대략 총 63억 원의 예산과 약 11,000명 이상의 인원이 동원된 것이다.

지 않는다. 또한 국어 교과의 서답형 답안에서는 문법이나 문학 영역의 용어처럼 고유한 교과 용어를 사용하면서도 다른 일상용어를 아우르기도 한다.

둘째, 동일한 용어를 사용하더라도 그 용례가 각 교과마다 다르다. 사회 교과의 '수요'와 '공급', '비교 우위' 등의 용어는 일상적으로도 사용될 수 있으나 사용 맥락이나 그 의미에서 차이가 난다. 마찬가지로 과학 교과에서 '녹는다'라는 의미는 일상에서 사용되는 경우보다 더 엄격한 개념 규정을 지닌다. 마찬가지로 국어 교과에서 일반적으로 유사어로 인정하지만 분야에 따라 인정하지 않는 경우와 그 역의 경우가 존재한다. 가령, 사회 교과에서 수요와 수요량은 다르며, 과학 교과에서 속력과 속도는 다른 개념이므로 채점 시 이를 명확히 구분하여 점수를 부여하는데, 국어 교과에서는 이들을 유사 의미로 처리할 수도 있다.

셋째, 각 교과마다 출제되는 문항의 형식이 다양하다. 예를 들어 국어 교과에서는 제시되는 지문의 글을 읽고 이해한 다음, 명사형 단어 외에도 술어형 단어나 구, 문장 형태의 답안을 다양하게 요구한다. 사회 교과에서도 자료 해석에 따른 문장형 답안을 요구하는 문항도 출제되고 있으나 다른 교과에 비해 술어형 답안보다는 사회과학 용어나 내용함축적 개념어 형태의 답안을 요구하는 문항이 자주 보인다.

이와 같이, 교과마다 주 사용 용어 데이터베이스가 다르고 문항 출제 형식이 다양하므로, 향후 자동채점 프로그램의 지속적 발전을 위해서는 교과 간 문항의 특성 및 자동채점 결과의 비교가 필요하다. 이에 본 연구에서는 2012년 국가수준 학업성취도 평가(이하 학업성취도 평가)의 초·중·고 국어와 초·중 사회/과학의 총 38문항 각 3,010개 답안을 대상으로 하여, 먼저 교과 간 문항 및 답안 유형의 차이를 살펴보고(Ⅲ장), 다음으로 교과 간 자동채점 결과의 차이를 비교해 보고자 한다(Ⅳ장). 자동채점 결과는 자동채점의 채점 비율은 물론, 표집 채점을 통해 확정된 기준 점수와 자동채점 점수 간 일치도(Kappa계수 및 상관계수)와 채점 불일치 비율을 분석한다.

이러한 교과 간 답안 유형 및 자동채점 결과의 비교를 통해 각 교과의 특성이 어떠한지 검토하는 것은 자동채점의 정확성 및 효율성 향상을 꾀할 수 있으므로, 향후 대규모 평가를 위한 서답형 자동채점 프로그램의 발전 방안을 모색하는 데 도움을 줄 수 있을 것으로 기대한다.

Ⅱ. 자동채점 프로그램 선행 연구

오늘날 세계 각국과 국제적 규모의 학업성취도 평가에서 컴퓨터 기반 평가 체제의 도입을 적극적으로 모색하고 있다. 실제로 미국의 GMAT(Graduate Management Admission Test)와 NAEP(National Assessment of Educational Progress), 네덜란드의 MATHCAT(대

학의 수학 강의 초기 배치평가), 일본의 CASEC(Computerized Assessment System for English Communication) 등의 컴퓨터 기반 평가가 실시되고 있으며, PISA(Programme for International Student Assesment)와 ICILS(International Computer and Information Literacy Study)와 같은 국제 학업성취도 평가에서도 컴퓨터 기반 평가를 도입·시행하고 있다(김경희 외, 2013, pp. 1-3). 이에 우리나라에서도 국제적인 평가 동향에 발맞추어 향후 국가수준 학업성취도 평가에서 컴퓨터 기반 평가 시스템을 도입을 고려하고 있다.[3)]

향후 대규모 평가들이 컴퓨터 기반으로 시행된다면 서답형 문항의 자동채점 프로그램은 필연적으로 요구되는 시스템이다. 컴퓨터 기반 평가는 학생들 답안을 빠르게 수합하고 채점하여 그 결과를 신속히 송환하는 것이 큰 장점이므로, 이를 극대화하기 위해서는 무엇보다 기계에 의한 자동채점이 필수적으로 실현되어야 한다. 현재의 학업성취도 평가와 같이 교과 지식 기반으로 답안 내용이 제한적이고 그 언어 단위 길이가 짧을수록 기계에 의한 자동채점 가능성과 채점 신뢰도는 높아지며, 특히 대규모일수록 예산 절감 효과가 커진다. 따라서 우선 현재의 자연언어처리 기술로 해결 가능한 단어·구 수준의 짧은 답안부터라도 자동으로 채점할 수 있는 프로그램을 개발하고 이를 적용하는 연구를 시도할 필요가 있다.

서답형 문항 자동채점 프로그램은 컴퓨터 기술을 이용해서 서술식으로 기술된 답안을 자동으로 평가하고 채점하는 것으로(Dikli, 2006, p. 4), 크게는 에세이형 채점 프로그램과 단문형 채점 프로그램으로 나뉜다. 에세이형 채점 프로그램은 보통 답안에 포함된 내용 자체의 정확성보다 글 쓰는 방식에 초점을 두는 반면, 단문형 채점 프로그램은 자연언어처리 기반으로 내용의 정확성을 중심으로 채점한다(Butcher & Jordan, 2010, p. 489).

영어의 경우에는 1960년대에 자동채점 프로그램 개발이 시작되어 현재 상용하고 있으며 다른 언어의 경우에도 최근에 점차 연구·개발하고 있다(Shermis & Burstein, 2003, p. xi). 영어 자동채점 프로그램의 경우는 약 60년 이상의 역사를 가지고 대규모 평가에 대비하기 위해 꾸준하게 연구·개발되어 왔으며, 대규모 평가에서는 채점 보조 수단[4)]으로 이용되거나 그 활용 가능성을 예비로 탐색[5)]하고 있는 중이다. 대부분의 영어 자동채점 프로그램이 에세이형 작문

3) 교육부는 '스마트교육 추진 전략(2011. 6. 29)'에 따라 온라인 평가 체제를 구축하여 학업성취도 평가를 IBT 방식으로 전환하려는 계획을 발표하였다. 이러한 계획을 실현하기 위해 김경희 외(2013)의 연구에서는 '2013년 기본계획 수립 ⇒ 2014~2015년 시범 적용 ⇒ 2016~2017년 확대 적용 ⇒ 2018~2019년 안정화'와 같이 컴퓨터 기반 국가수준 학업성취도 평가의 단계별 도입 방안을 제안하였다(김경희 외, 2013, p. 8).

4) 자동채점 프로그램을 채점 보조 수단으로 사용하는 대표적인 평가 도구로는 GMAT(Graduate Management Admission Test, 미국 경영대학원 입학시험), TOEFL(Test of English as a Foreign Language), GRE(Graduate Record Examination), 메릴랜드 주 과학 시험, 미국의 대학 입학 후 컴퓨터 기반 학력 진단시험인 ACCUPLACER 시험, Pearson 영어 시험 등을 들 수 있다(노은희 외, 2012, p. 29).

5) 미국의 경우, 1990년대 초부터 교육과정을 구현하고 교육의 변화를 유도하기 위해서 서답형 문항의 활용이 필요하다는 판단 하에, NEAP와 주 단위의 대규모 검사에서 서답형 문항을 도입하기 위한 채점 방안 연구를 진행하였다(Ben-Simon & Bennett, 2007; NAEP, 2010; Attali, 2011). 대규모 평

능력을 채점하는 데 주목적이 있으나, ETS에서 개발한 C-rater는 비교적 짧은 서답형 답안에 대해 개념 중심으로 답안을 평가하고 채점하는 시스템이다. 현재 한국의 대규모 평가에서 활용되는 서답형 문항은 한 교과 영역에서 주요 개념의 이해 정도를 평가하기 위한 단문 형태의 답안을 요구하므로, 우선적으로 단어·구 수준의 한국어 서답형 문항 자동채점 프로그램을 개발하는 데 참조할 만하다(노은희 외, 2012, pp. 31-37).

그럼에도 불구하고, 프로그램 내의 알고리즘은 일종의 원천 기술이기 때문에 공개하고 있지 않으며 영어가 아닌 한국어로 적용하는 것은 한국어 특징에 부합하는 자연언어처리 기술을 요구하므로, 이들 프로그램을 따라 개발하는 데는 제한점이 있다(노은희 외, 2012, p. 39). 사실, 한국어의 경우 지금까지 자동채점화 방안을 본격적으로 논의하지 못하고 있는 실정이다. 한국어 자동채점 프로그램 개발과 관련하여 특정 영역의 소규모 시험을 대상으로 실험실 수준의 몇몇 시도는 있었으나(〈표 1〉 참조), 실험 결과가 타당도와 신뢰도 측면에서 검증되지 않았고 대규모 평가나 다른 영역에서 활용 가능한지는 추후 논의가 필요한 상태이다. 이와 같은 한국어 자동채점 프로그램의 경우 개인 연구 차원에서 시도되는 초보적인 개발 단계에 머물러 있어서 대규모 평가에 적용하기에는 여러 한계를 지니고 있다.

〈표 1〉 한국어 서답형 문항 자동채점 프로그램 개발 연구

연구자 (연도)	프로그램 특성	한계
정동경 (2001)	- 벡터 유사도와 시소러스를 이용하여 채점 - 핵심어의 일치 정도를 이용해서 채점하는 비교적 단순한 방법을 활용함.	- 부정문 등으로 완전히 다른 의미로 표현된 답안의 경우 핵심어만 일치하면 정답으로 채점할 수 있음. - 부분 점수를 고려하지 않음.
박희정 강원석 (2003)	- 문항 유형을 제안하고 유의어 사전과 정보검색 개념을 이용하여 채점 - 휴리스틱 유사도 계산 알고리즘을 사용해서 학생 답안에 불필요한 단어가 지나치게 많이 나타났을 경우를 반영하여 긴 답안이 무조건 좋은 점수를 받을 수 없도록 함.	- 부정문 등으로 완전히 다른 의미로 표현된 답안의 경우도 핵심어만 일치하면 정답으로 채점할 수 있음. - 문장이 가진 의미보다는 출현 단어의 일치도만 검사한다는 한계가 있음.
권오영 (2004)	- 문항의 유형과 가능한 정답의 유형에 따라 모범 답안을 간단한 논리식 형식으로 기술하고 학생 답안이 이 논리식에 부합하는지를 검사하도록 설계함. - 논리식에 정확하게 일치하는 경우는 정확하게 채점할 수 있음.	- 이 프로그램을 적용하기 위해서는 정의한 논리식에 따라 문항을 출제하고 모범 답안을 작성해야 함. - 정형화된 문장이 아닌 다양한 형태의 문장으로 확장할 때는 제약이 따름.

가에서 서답형 문항 관련 연구는 대부분 채점 방법과 관련되며, 최근에 서답형 평가의 자동채점 방안이 활발히 적용·연구되고 있다(노은희 외, 2012, p. 38).

연구자 (연도)	프로그램 특성	한계
조우진 (2006)	- 정보검색에서 사용되는 의미 커널(semantic kernel)과 한글 워드넷(Korean WordNet)을 이용해서 학생 답안을 평가하고 채점 - 이 프로그램은 비슷한 의미로 작성된 문장에 대해서 동일한 채점을 할 수 있다는 장점이 있음.	- 모범 답안 집합이 필요함. - 부정문 등으로 완전히 다른 의미로 표현된 답안의 경우 정답으로 채점할 수 있음.
강원석 (2011)	- 문항의 질의문 유형에 따라 채점 방법을 다르게 적용함. - 규칙을 이용해서 질의 유형에 따라 다른 채점 방법을 이용함으로써 좀 더 신뢰성 있는 채점을 할 수 있음.	- 인간채점과 자동채점의 상관계수만 비교하여 채점의 정확도는 확인할 수 없음. - 형태소 분석 정도만 이용하고 있어서 부정문 등으로 완전히 다른 의미로 표현된 답안도 정답으로 채점할 가능성 있음.

출처: 노은희 외, 2012, p. 40

이러한 상황에서 한국교육과정평가원은 대규모 평가를 염두에 두고 2006년부터 서답형 문항의 자동채점 연구에 관심을 기울여 왔다(〈표 2〉 참조). 2006년 대규모 자동채점 시스템 도입 방안에 대한 3개년도 연구를 시작으로(진경애 외, 2006; 2007; 2008), 2010년 학업성취도에 쓰인 서답형 문항 중 컴퓨터 자동채점이 가능한 문항을 선별하고 이를 모의 적용한 바 있으나 실제 프로그램 개발로 이어지지는 못했다(성태제 외, 2010). 2012년 초·중·고 학생들의 영어 말하기와 쓰기 인증 시험의 대규모 채점 부담을 해소하기 위해 자동채점 프로그램을 개발하였으며(신동광 외, 2012; 시기자 외, 2012), 한국어 서답형 문항에 대해서도 단어·구 수준에서 자동채점할 수 있는 프로그램을 개발·보완하고 그 채점 신뢰도를 검증하는 연구를 진행하였다(노은희 외, 2012; 2013).

〈표 2〉 한국교육과정평가원의 서답형 문항 자동채점 관련 연구

연구자(연구년도)	연구명	연구 목적
진경애 외 (2006, 2007, 2008)	서답형 문항 자동채점 시스템 도입 방안 연구 Ⅰ, Ⅱ, Ⅲ	영어 서답형 문항 자동채점 프로그램 도입 가능성 탐색 및 영작문 자동채점 프로그램 개발
성태제 외 (2010)	학업성취도평가 서답형 문항 컴퓨터 채점화 방안 탐색	한국어 서답형 문항의 컴퓨터화 채점 방안 탐색 및 실현 가능성 확인
신동광 외 (2012)	국가영어능력평가시험의 말하기 자동채점 프로그램 도입 방안	영어 말하기 자동채점 프로그램 프로토타입 개발
시기자 외 (2012)	국가영어능력평가시험 쓰기 자동채점 프로그램 개발	영어 쓰기 자동채점 프로그램 프로토타입 개발
노은희 외 (2012)	대규모 평가를 위한 서답형 문항 자동채점 방안 연구	한국어 단어·구 수준 서답형 문항 자동채점 프로그램 프로토타입 설계·개발
노은희 외 (2013)	대규모 평가를 위한 서답형 문항 자동채점 프로그램 정교화 및 시범 적용	기 개발된 한국어 단어·구 수준 서답형 문항 자동채점 프로그램 정교화 및 적용 가능성 제고 방안 수립

Ⅲ. 교과별 문항 및 답안 유형 특징 분석

1. 2009~2012년 학업성취도 평가 교과별 서답형 문항 현황

본 연구의 적용 대상인 2012년 학업성취도 평가 국어, 사회, 과학 문항에 대한 교과별 세부 분석에 앞서, 전수평가가 안정적으로 정착된 2009~2012년 최근 4년 동안의 교과별 서답형 문항 현황을 살펴보면 〈표 3〉과 같다.

〈표 3〉 교과별 서답형 문항의 답안 유형별 문항 수 및 비율 (2009~2012년)

구분		소문항 수* (문항 수)	소문항의 답안 유형별 문항 수 및 비율							
			단어·구 답안		문장 답안		다문장 답안		기타 답안**	
			문항 수	비율	문항 수	비율	문항 수	비율	문항 수	비율
2009	국어	63 (28문항)	39	61.9	15	23.8	9	14.3	0	0
	사회	28 (18문항)	26	92.9	1	3.6	1	3.6	0	0
	과학	41 (14문항)	36	87.8	2	4.9	0	0	3	7.3
2010	국어	34 (17문항)	22	64.7	5	14.7	7	20.6	0	0
	사회	18 (15문항)	15	83.3	3	16.7	0	0	0	0
	과학	25 (12문항)	22	88	1	4	1	4	1	4
2011	국어	32 (17문항)	22	68.8	4	12.5	6	18.8	0	0
	사회	19 (15문항)	15	78.9	1	5.3	3	15.8	0	0
	과학	32 (12문항)	23	71.9	3	9.4	3	9.4	3	9.4
2012	국어	36 (17문항)	26	72.2	4	11.1	6	16.7	0	0
	사회	24 (15문항)	21	87.5	1	4.2	1	4.2	1	4.2
	과학	23 (12문항)	13	56.5	5	21.7	4	17.4	1	4.3
전체	국어	165 (79문항)	109	66.1	28	17.0	28	17.0	0	0
	사회	89 (63문항)	77	86.5	6	6.7	5	5.6	1	1.1
	과학	121 (50문항)	94	77.7	11	9.1	8	6.6	8	6.6
	계	375(192문항)	280	74.7	45	12	41	10.9	9	2.4

* 소문항 수는 자동채점 프로그램 적용을 위하여 하위 문항을 소문항으로 분류하여 계산함.
** 기타 답안은 그래프나 그림, 수식 등을 요구하는 답안을 의미함.

〈표 3〉에서 세 교과를 종합하면, 2009~2012년 학업성취도 평가의 서답형 문항에서 답안의 유형별 비율은 소문항 기준 총 375문항 중, 단어·구 답안이 280문항(74.7%)으로 가장 높았

고, 문장 답안이 45문항(12%), 다문장 답안이 41문항(10.9%), 기타 답안이 9문항(2.4%)을 차지하였다.

교과별로 살펴보면, 단어·구 답안의 경우 사회 교과가 86.5%로 가장 높았으며, 과학 교과가 77.7%, 국어 교과가 66.1% 순으로 나타났다. 반면, 문장 답안과 다문장 답안은 국어 교과가 각각 17.0%로 가장 높았으며, 과학 교과가 각각 9.1%, 6.6%, 사회 교과가 각각 6.7%, 5.6% 순이었다. 기타 답안의 경우 과학 교과가 6.6%로 상대적으로 비율이 매우 높았으며, 국어 교과(0%)와 사회 교과(1.1%)는 거의 출제되지 않았다. 이는 앞서 언급한 바와 같이, 교과간 문항 출제 형식의 차이에 따라 술어형 단어나 구, 문장 형태의 답안을 주로 요구하는 문항을 출제하는 교과가 있는 반면 내용함축적 개념어 형태의 답안을 요구하는 문항을 출제하는 교과가 있기 때문인 것으로 추정된다. 즉, 서답형 문항에서 요구하는 교과 간 언어 단위의 차이는 한국어 답안을 처리하는 자연언어처리 기술과 그에 따른 자동채점 양상이 교과마다 다를 수 있음을 짐작케 한다.

2. 정답 패턴 구분 기준

본 연구의 개발 프로그램은 단어·구 수준 답안의 처리를 일차 목표로 한다. 따라서 현재의 자연언어처리 기술 및 개발 프로그램이 처리 가능한 서답형 답안을 판정하기 위해서는 문항이 요구하는 답안에 대한 정밀한 분석이 필요하다.

이에 노은희 외(2012)의 연구에서는 한국어 서답형 답안을 대상으로 문자열 일치(스트링 매치), 형태소 분석, 구문 분석, 의미/담화 분석 등의 자연언어처리 기술 필요 여부에 따라 정답 패턴을 P1에서 P6까지 구분하고, 기타로 그래프, 선긋기, 체크박스로 답안을 작성하는 경우는 각각 P7, P8, P9로 구분하였다. 본 연구도 이러한 정답 패턴에 기초하되, P1~P6을 다음과 같이 보다 구체화하여 살펴보고자 한다.

〈표 4〉 한국어 서답형 문항의 정답 패턴 분석과 예시

구분	P1	P2	P3	P4	P5	P6
자연언어처리 기술*	문자열 일치	형태소 분석 (어휘 토큰 분석)	형태소 분석 (어휘·문법 토큰 분석)	단문 구문 분석	복문 구문 분석	의미 분석, 담화 분석
적용 언어 단위	단어	단어·구	구	단문의 문장	단문 및 복문의 문장	복문 및 다문장
사용 용어** 수	1~2	2~3	3~4	4~6	6~8	8~10

구분	P1	P2	P3	P4	P5	P6
예시	북극곰	북극곰의 눈물	지구온난화로 빙하가 녹아서	지구온난화로 남극과 북극의 빙하가 녹고 있다.	지구온난화로 북극의 빙하가 녹아서 북극곰들이 눈물을 흘리다.	지구온난화는 북극의 빙하를 녹인다. 북극의 빙하가 녹으면 해수면이 상승한다. 그래서 태평양의 여러 섬들이 잠기게 된다.

* 자연언어처리 기술이 병렬적으로 표현되어 있으나, P2부터 이전 정답 패턴에 사용된 자연언어처리 기술이 누적되어 사용됨을 의미함.

** 한국어 조사, 어미를 제외하고 띄어쓰기를 하는 주요 내용어를 의미함.

〈표 4〉에서 P1은 형태소 분석이 필요 없이 정답과의 일치 여부를 판정하면 되는 경우이다. P2는 형태소 분석 중 어휘 토큰[6] 분석이 필요한 것이며, 단순히 부분 문자열만 일치하면 정답으로 간주하는 유형이다. P3은 형태소 분석 중 어휘와 문법 토큰 분석이 필요한 경우로, 형태소 분석을 통해 추출된 토큰들에 대해서 어휘형태소뿐만 아니라 문법형태소(조사 또는 어미) 부분까지 반영하여 정답 여부를 판정해야 하는 유형이다. P4와 P5는 구문 분석을 통해 개념이 동일한지를 판단해야 하는 경우로 P4는 단순 개념의 구문 분석, P5는 복합 개념의 구문 분석이 요구된다. P6의 경우에는 복문 및 다문장 수준으로 서술하는 문항으로 답안을 채점하기 위해 의미와 담화 분석이 요구되기도 한다.

3. 교과별 적용 대상 문항 분석

2013년에 개발한 한국어 서답형 문항 자동채점 프로그램은 일차적으로 단어·구 수준 답안 처리를 목표로 하므로, 본 연구의 시범 적용 대상은 주로 단어·구 수준 답안으로 한정하고자 한다. 〈표 5〉는 본 연구의 자동채점 적용 대상인 2012년 학업성취도 평가 국어, 사회, 과학 문항에 대한 정보를 정리하여 나타낸 것이다.

6) 자연언어처리에서 문장(또는 문자열, string)을 미리 정의되어 있는 최소 단위의 문자열들로 분할했을 때 분리된 각 부분 문자열(substring)을 '토큰'이라고 한다.

〈표 5〉 2012년 국어, 사회, 과학 자동채점 대상 서답형 문항 정보

교과	학교급	문항 번호	정답 패턴	답안 유형 수 /1000	답안 작성 유형	모범 정답
국어	초6(5문항)	2-(1)	P1	10	1단어 명사형	송편
		2-(2)	P1	10	1단어 명사형	곡식
		2-(3)	P1	15	1단어 명사형	무
		5-(1)	P2	55	1단어 술어형	밝게(환하게, 환히)
		5-(2)	P2	46	1단어 술어형	크게(커)
	중3(4문항)	1-(1)	P2	107	1단어 술어형	타당(적절)
		1-(2)	P2	62	1단어 명사형	근거
		5-(1)	P3	7	1단어 술어형	아름이를 생포하였다.
		5-(2)	P1	53	1단어 명사형	아름이
	고2(8문항)	1-(1)	P1	6	1단어 명사형	대조
		1-(2)	P1	8	1단어 명사형	강조
		4-(1)-ⓐ	P1	12	1단어 명사형	㉠
		4-(1)-ⓑ	P2	69	1단어 술어형	알맞은
		4-(2)	P4	311	3단어 술어형	사회 구조가 복잡하고
		5-(1)	P1	30	1단어 술어형	높이
		5-(2)	P1	33	1단어 술어형	깊이
		5-(3)	P1	90	1단어 술어형	평평한
사회	초6(6문항)	1	P1	59	1단어 명사형	우데기
		3	P1	57	1단어 명사형	서학
		4-(1)	P1	39	1단어 명사형	양반
		4-(2)	P1	40	1단어 명사형	천민
		6-(1)	P3	171	1단어 술어형	증가하였다.
		6-(2)	P1	40	1단어 명사형	고령화
	중3(7문항)	3-(1)	P1	83	1단어 명사형	도심
		3-(2)	P1	155	1단어 명사형	부도심
		5-(1)	P1	162	1단어 명사형	사헌부
		5-(2)	P1	152	1단어 명사형	의금부
		6-(1)	P4	201	2단어 술어형	신항로 개척
		6-(2)	P3	97	1단어 술어형	감소하였고
		6-(3)	P3	114	1단어 술어형	상승하였다.
과학	초6(4문항)	2-(1)-ⓐ	P1	31	1단어 명사형	양
		2-(1)-ⓑ	P3	129	2단어 명사형	녹는 양
		3-(1)	P1	43	1단어 명사형	콩팥(신장)
		3-(2)	P5	594	1문장	노폐물을 거른다.
	중3(4문항)	5-(1)	P1	45	1단어 명사형	대뇌
		6-(1)	P1	112	1단어 명사형	뿌리털
		7-(1)	P1	29	기호배열	(나)-(다)-(가)
		7-(2)	P2	202	2단어 명사형	습곡 산맥

소문항 수를 기준으로 할 때, 국어 문항은 17문항, 사회 문항은 13문항, 과학 문항은 8문항으로 총 분석 대상의 문항 수는 38문항이다. 정답 패턴별로 살펴보면 P1 유형이 총 23문항으로, 전체 38문항 중 60.5%의 높은 비율을 차지하며, P2 유형과 P3 유형이 각각 7문항, 5문항이며 P4 유형과 P5 유형은 각각 2문항, 1문항이다. 이 가운데 P1~P3 유형의 총 35문항은 본 연구의 분석 대상인 단어·구 답안에 해당하며 이는 형태소 분석 기술로도 자동채점이 가능하다. 현재 한국어 자연언어처리 기술 중에서 형태소 분석 기술은 비교적 정교하게 마련되어 있는 편이어서 단어·구 답안을 처리하는 데는 큰 장애가 없을 것으로 판단된다.

정답 패턴 유형의 숫자가 커질수록 자연언어처리 기술이 점차 더 요구됨을 의미하고, 이는 곧 자동채점 가능성이 그에 따라 낮아질 수 있음을 뜻한다. 교과별로 볼 때, 국어 교과는 총 17개 문항 중 P1 문항이 10개, P2 문항이 5개, P3 문항이 1개, P4 문항이 1개가 분석 대상에 포함되었으며, 사회 교과는 총 13개 문항 중 P1 문항이 9개, P2 문항이 0개, P3 문항이 3개, P4 문항이 1개가 분석 대상에 포함되었다. 과학 교과는 총 8개 문항 중 P1 문항이 5개, P2 문항이 1개, P3 문항이 1개, P5 문항이 1개가 분석 대상에 포함되었다. 여기서 국어 교과의 경우, 사회와 과학 교과에 비해 자동채점 처리가 용이한 P1~P2 문항 비중이 높아 자동채점 비율이 높을 것으로 예상된다.

한편, P4 유형과 P5 유형의 3문항은 문장 수준 답안이거나 구문 분석 기술까지 요구하는 답안에 해당하는 것으로 향후 프로그램 개발을 위한 점검 차원에서 실험적으로 자동채점 대상 문항으로 포함시켰다. 가령, P4 유형인 고등학교 국어 4-(2)의 '사회 구조가 복잡하고'의 경우 모범 답안은 3단어 술어형으로 간단해 보여도, 학생 답안은 '사회구조가 복잡해질수록 그리고', '복잡해지는 사회구조 속에', '사회구조가 상호보완적이고 사회구조가 복잡하고' 등으로 다양하게 나타나 이에 대한 점수 인정 여부를 판단하려면 복잡한 의미 분석 처리 기술이 요구된다. 마찬가지로 P5 유형인 과학 중학교 3-(2)의 경우도 모범 답안은 '노폐물을 거른다'이나 실제 문항에서는 언어 형식의 제약 없이 콩팥의 기능을 묻는 것이어서, 유사 답안으로 '노폐물이 줄어든다', '오줌을 만든다', '혈액이 깨끗해진다'는 의미가 포함된 경우 모두 정답으로 처리한다. 실제 학생들의 답안도 '노폐물을 제거하고 혈액을 정화시킨다', '신장이 오줌을 만들어서 혈액 속에 있는 노폐물을 빼주는 역할을 한다', '피를 정화시켜 노폐물을 걸러줘서 피를 맑게 만들어 준다'와 같이 여러 다른 정보를 보태어 정답을 늘여 쓴 경우도 많다.

단어·구 수준인 P1~P3 문항에 대해 답안 1,000개를 기준으로 하여 서로 다른 답안의 유형 수를 살펴보면, 국어 문항은 6~107개 사이로 학생 답안 간 편차가 비교적 작게 나타났다. 반면, 사회와 과학 문항은 각각 39~171, 29~202개 사이로 학생 답안 간 편차가 국어 문항에 비해 상대적으로 크게 나타났다. 이는 국어 문항에 비해 사회, 과학 문항에서 학생들이 더 다양하게 응답하고 있음을 의미한다. 국어 문항의 경우 다른 교과에 비해 문장이나 다문장 수준 문항의 비율이 높아 학생들의 다양한 답안 반응은 주로 이러한 문항에서 요구하는 편이고, 단어·구 수준 문항에서는 언어 형식적 제약을 통해 답안의 다양성을 〈보기〉나 〈조건〉을 통해 제약

한다. 이에 비해 사회, 과학 교과와 같이 내용 기반의 교과에서는 학생 답안이 다양하더라도 관련 개념이나 의미로 제약이 어느 정도 가능하기 때문에 크게 언어 형식으로 제약하지 않기 때문이다.

자동채점의 가능성은 정답 패턴, 답안 유형 수, 답안 작성 유형을 복합적으로 고려하여 결정된다. 이 중, 답안 유형 수를 기준으로 현재 개발된 프로그램의 자동채점 가능성을 살필 때, 대개 답안이 단어・구이면서 답안 유형 수가 1,000개 당 100개 미만인 경우 자동채점 가능성이 '매우 높음'으로, 100~200개인 경우 자동채점 가능성이 '높음으로', 200~400개인 경우 '보통'으로, 400~600개인 경우 '낮음'으로, 600개 이상인 경우 '매우 낮음'으로 분류하였다(노은희 외, 2012, p. 52). 이를 통해 볼 때, 단어・구 수준인 P1~P3 문항에서 국어 교과는 16개 문항 중 100개 미만인 문항이 15개여서 대체로 분석 문항의 자동채점 가능성이 '매우 높음'으로 나타난다. 사회 교과는 12개 문항 중 100개 미만인 문항이 7개이고, 100~200개인 문항이 5개여서 자동채점 가능성이 '매우 높음' 또는 '높음'으로 나타나나 국어 교과보다는 그 가능성이 낮아진다. 과학 교과는 7개 문항 중 100개 미만인 문항이 4개이고, 100~200개인 문항이 2개, 200~400개인 문항이 1개여서 자동채점 가능성이 '매우 높음' 또는 '높음'이 대체적이나 '보통'인 문항도 있어서 국어와 사회 교과보다는 자동채점이 까다로울 것으로 예상된다.

Ⅳ. 자동채점 결과 분석

교과 간 자동채점 결과를 비교하기 위해, 2012년 국가수준 학업성취도 평가 국어 초・중・고 17문항, 사회 초・중 13문항, 과학 초・중 8문항의 각 3,010개 학생 답안을 자동채점 프로그램을 사용하여 채점하였다. 자동채점 결과는 자동채점의 채점 비율, 표집 채점을 통해 확정된 기준 점수와 자동채점 점수 간 일치도(Kappa계수 및 상관계수), 채점 불일치 비율을 분석하였다. 채점자 간 신뢰도를 추정하는 방법은 여러 가지가 있으나 자동채점 신뢰도 연구에서는 일치도 통계(agreement statistics), Kappa계수, 상관계수 등이 주로 사용된다(성태제, 2014, p. 454). 일치도 통계는 계산하기 쉽고 설명하기 용이하다는 장점이 있으나, 우연에 의해 동일하게 분류되는 비율을 포함하므로 과대 추정되는 경향이 있어서, 두 채점자의 채점 결과에서 우연에 의해 일치되는 비율을 제외하고 계산하는 Cohen′s Kappa계수를 사용하기도 한다(성태제, 2014, p. 460). Kappa계수는 우연에 의한 일치 비율을 제거하므로 일치도 통계보다 더 정확하다고 할 수 있다(Meadows & Billington, 2005, p. 14). 본 연구의 적용 대상인 학업성취도 평가 서답형 문항은 정답이 명확한 단어・구 수준이고, 채점 점수가 범주화되어 있으며 채점자 간 점수 범주의 일치도를 파악하고자 하므로 Kappa계수를 사용하는 것이 정확하다. 다

만, 많은 자동채점 연구에서는 상관계수와 Kappa계수를 동시에 제시하고 있으므로 본 연구에서도 Kappa계수와 더불어 참고용으로 상관계수를 함께 제시한다.

1. 국어 문항 자동채점 결과

먼저, 국어 문항의 학교급별 Kappa계수 및 상관계수,[7] 채점 불일치 비율[8]을 살펴보면 〈표 6〉과 같다. 국어 문항에서 P1, P2에 해당하는 서답형 문항의 Kappa계수와 상관계수는 최소 .98 이상이며 절반 이상의 문항이 1.00으로 채점 신뢰도가 매우 높다[9]고 할 수 있다. 또한 이들 문항은 채점 불일치 비율이 0~0.2%로 매우 낮게 나타났다.

〈표 6〉 2012년 국어 문항 인간채점과 자동채점 간 Kappa 및 상관계수, 채점 불일치 비율

학교급	문항 번호	정답 패턴	사례 수	Kappa 계수	상관 계수	채점 불일치 비율	모범 정답
초6	2-(1)	P1	3,010	.98	.98	0.07	송편
	2-(2)	P1	3,010	.98	.98	0.07	곡식
	2-(3)	P1	3,010	1.00	1.00	0.03	무
	5-(1)	P2	3,010	.99	.99	0.23	밝게(환하게, 환히)
	5-(2)	P2	3,010	.99	.99	0.10	크게(커)

7) 일치도의 기준 점수로 사용되는 인간채점 점수는 사전에 표집 채점에서 2인의 복수채점자가 3라운드에 걸쳐 해당 답안에 대해 확정한 점수를 사용한다.

8) 채점 불일치 비율 = 채점 불일치 수 / (전체 답안 수 - 미판단 답안 수) × 100

9) 성태제(2014, p. 462)는 채점자 간 신뢰도를 판단하는 절대적인 기준은 없으나 채점 결과가 점수로 제공될 때는 .6 이상, 채점 결과가 등급이나 범주로 제시될 때는 일치도 계수 .85 이상(Kappa계수는 .75 이상)을 제안하였다. Nehm 외(2012)는 Landis와 Koch(1977, p. 159)의 의견을 참조하여, Kappa계수가 .41~.60이면 '적정'하고 .61~.80이면 '실질'적이며, .81~1.00이면 '거의 완전'하다고 평가하였다(Nehm, Ha, & Mayfield, 2012, p. 187).
자동채점 분야에서 채점 신뢰도는 문항 특성, 채점 기준이나 평가 상황 등에 따라 다르게 해석해야 할 것이다. 에세이와 같이 응답의 자유도가 높고 총괄적 채점 방법을 사용하는 경우 또는 채점자의 주관성이 작용하는 경우에는 채점자 간 신뢰도가 일반적으로 낮고, 정답이 명확한 서답형 문항의 경우에는 채점자 간 신뢰도가 높다. 또한 평가 상황이 교수·학습 상황처럼 고부담이 아닌 경우에는 채점자 간 신뢰도의 기준이 엄격하지 않을 수 있으나, 평가 결과가 의사결정에 활용되는 고부담 시험의 경우에는 채점의 정확도가 매우 중요하므로 엄격한 기준 적용이 요구된다. 본 연구에서 자동채점 프로그램에 적용하고자 하는 학업성취도 평가는 대규모 고부담 시험이므로 엄격한 신뢰도 기준이 요구된다. 또한 적용 대상 문항의 평가 내용이 비교적 명확하고 제한되어 있는 단어·구 수준이므로 높은 일치도를 보인다. 따라서 본 연구에서는 인간채점과 자동채점 간 Kappa계수가 .80 이상인 문항을 신뢰할 만하다고 해석하고자 한다.

학교급	문항 번호	정답 패턴	사례 수	Kappa 계수	상관 계수	채점 불일치 비율	모범 정답
중3	1-(1)	P2	3,010	.99	.99	0.56	타당(적절)
	1-(2)	P2	3,010	.99	.99	0.50	근거
	5-(1)	P3	3,010	.97	.97	0.34	아름이를 생포하였다.
	5-(2)	P1	3,010	.99	.99	0.20	아름이
고2	1-(1)	P1	3,010	1.00	1.00	0.03	대조
	1-(2)	P1	3,010	1.00	1.00	0	강조
	4-(1)-ⓐ	P1	3,010	1.00	1.00	0.23	㉠
	4-(1)-ⓑ	P2	3,010	1.00	1.00	0.20	알맞은
	4-(2)	P4	3,010	.98	.99	1.02	사회 구조가 복잡하고
	5-(1)	P1	3,010	1.00	1.00	0.10	높이
	5-(2)	P1	3,010	1.00	1.00	0.17	깊이
	5-(3)	P1	3,010	1.00	1.00	0.20	평평한

* 소수점 셋째 자리에서 반올림

다만 중3 1-(1)번, 1-(2)번, 고2 4-(1)-ⓐ번 문항의 경우 채점 불일치 비율이 각각 0.56, 0.50, 0.23%로 다른 P1, P2 문항에 비해 다소 높게 나타났는데, 이는 채점자의 채점 오류가 발생하였거나 학생 답안에 정답 외의 다른 단어·구가 함께 쓰인 경우 자동채점 프로그램이 정답으로 처리하였기 때문이다. 전자의 경우 자동채점 프로그램이 인간 채점의 오류를 잡아줄 수 있다는 점에서 의의가 있으며, 후자의 경우 자동채점에서 정답 템플릿을 보다 정교하게 설계할 필요가 있음을 보여준다고 하겠다.

또한 3단어 술어형의 P3과 P4에 각각 해당하는 중3 5-(1)번 문항과 고2 4-(2)번 문항의 경우, Kappa계수가 .97과 .98로 나타나 다른 문항에 비해 상대적으로 낮았으며, 채점 불일치 비율은 0.34%, 1.02%로 나타나 다른 문항에 비해 상대적으로 높았다. 이를 통해 요구되는 답안이 길고 복잡해질수록 인간채점과 자동채점 결과의 일치도가 떨어지고 불일치 비율이 증가한다는 것을 알 수 있다. 그러나 검증된 모든 국어 문항의 Kappa계수와 상관계수가 .97이상으로 자동채점 결과는 타당하다고 하겠다.

2. 사회 문항 자동채점 결과

다음으로 사회 문항의 학교급별 Kappa계수 및 상관계수, 채점 불일치 비율은 〈표 7〉과 같다. 사회 문항에서 초6 3번 문항을 제외하고 P1에 해당하는 모든 문항이 Kappa계수와 상관계수가 1.00으로 채점 신뢰도가 매우 높고 채점 불일치 비율은 0~0.13% 정도로 매우 낮아 채점 결과가 타당한 것으로 나타났다. 초6 3번 문항 역시 Kappa계수와 상관계수가 .99로 채점

신뢰도가 매우 높다.

〈표 7〉 2012년 사회 문항 인간채점과 자동채점 간 Kappa 및 상관계수, 채점 불일치 비율

학교급	문항 번호	정답 패턴	사례 수	Kappa 계수	상관 계수	채점 불일치 비율	모범 정답
초6	1	P1	3,010	1.00	1.00	0.07	우데기
	3	P1	3,010	.99	.99	0.30	서학
	4-(1)	P1	3,010	1.00	1.00	0.07	양반
	4-(2)	P1	3,010	1.00	1.00	0	천민
	6-(1)	P3	3,010	1.00	1.00	0	증가하였다.
	6-(2)	P1	3,010	1.00	1.00	0	고령화
중3	3-(1)	P1	3,010	1.00	1.00	0.13	도심
	3-(2)	P1	3,010	1.00	1.00	0	부도심
	5-(1)	P1	3,010	1.00	1.00	0	사헌부
	5-(2)	P1	3,010	1.00	1.00	0	의금부
	6-(1)	P4	3,010	.98	.98	0.45	신항로 개척
	6-(2)	P3	3,010	1.00	1.00	0.07	감소하였고
	6-(3)	P3	3,010	1.00	1.00	0.07	상승하였다.

* 소수점 셋째 자리에서 반올림

또한 1단어 술어형의 P3에 해당하는 문항의 Kappa계수와 상관계수가 1.00으로, 2단어 술어형의 P4에 해당하는 문항의 Kappa계수와 상관계수가 .98로 나타나 사회 문항은 정답 패턴에 상관없이 대부분 채점 신뢰도가 높았다. 이는 사회 문항에서 요구하는 답안들이 '우데기', '사헌부', '고령화', '신항로 개척' 등과 같이 유의어가 비교적 적으면서도 명확한 개념어이거나 '증가하였다', '감소하였고' 등과 같이 유의어 사전 적용이 가능한 답안이기 때문인 것으로 추정된다. 그러나 국어나 과학에 비해 높은 채점 신뢰도에도 불구하고, P4에 해당하는 중3 6-(1)번 문항의 채점 불일치 비율이 0.45%로 P1, P3에 해당하는 사회과 다른 문항에 비해 높게 나타났다.

3. 과학 문항 자동채점 결과

마지막으로, 과학 문항의 학교급별 Kappa계수 및 상관계수, 채점 불일치 비율은 〈표 8〉과 같다. 과학 문항에서 P1, P2에 해당하는 서답형 문항의 Kappa계수와 상관계수는 중3 7-(2)를 제외하고 최소 .98 이상으로 나타나고 채점 불일치 비율이 0.07~0.17% 정도로 낮게 나타나 채점 신뢰도가 매우 높다고 할 수 있다.

〈표 8〉 2012년 과학 문항 인간채점과 자동채점 간 Kappa 및 상관계수, 채점 불일치 비율

학교급	문항 번호	정답 패턴	사례 수	Kappa 계수	상관 계수	채점 불일치 비율	모범 정답
초6	2-(1)-ⓐ	P2	3,010	.98	.98	0.17	양
	2-(1)-ⓑ	P3	3,010	.95	.95	2.10	녹는 양
	3-(1)	P1	3,010	.99	.99	0.13	콩팥(신장)
	3-(2)	P5	3,010	.64	.81	5.05	노폐물을 거른다.
중3	5-(1)	P1	3,010	1.00	1.00	0.10	대뇌
	6-(1)	P1	3,010	1.00	1.00	0.13	뿌리털
	7-(1)	P1	3,010	1.00	1.00	0.07	㈏-㈐-㈎
	7-(2)	P2	3,010	.96	.98	1.17	습곡 산맥

* 소수점 셋째 자리에서 반올림

다만 중3 7-(2) 문항의 경우 2단어 명사형으로 채점 기준에서 1단어 당 부분 점수를 부여하고 각 단어에 대한 유사답안이 다양하게 존재하기 때문에 Kappa계수와 상관계수가 P1, P2의 다른 문항들에 비해 상대적으로 낮게 나타났다. 또한 2단어 명사형의 P3에 해당하는 초6 2-(1)-ⓑ번 문항의 경우 Kappa계수와 상관계수가 .95로, 1문장형의 P5에 해당하는 초6 3-(2)번 문항의 경우 Kappa계수와 상관계수가 각각 .64와 .81로 나타났으며, 채점 불일치 비율은 초6 2-(1)-ⓑ번 문항이 2.10%, 초6 3-(2)번 문항이 5.05%로 높았다. 이를 통해 볼 때, 국어 문항과 마찬가지로 요구되는 답안이 길고 복잡해질수록 채점 신뢰도가 떨어지고 인간채점과 자동채점 결과의 불일치율이 증가한다는 것을 알 수 있다.

그러나 문장 수준 자동채점 프로그램 개발을 염두에 두고 실험적으로 검증한 P5의 초6 3-(2)번 문항을 제외하고 검증된 모든 과학 문항의 Kappa계수와 상관계수가 .95이상으로 자동채점 결과가 타당하다고 볼 수 있다.

4. 국어, 사회, 과학 문항 자동채점 결과 비교

국어, 사회, 과학 문항에 대한 자동채점 결과를 종합하면 〈표 9〉와 같다. 채점 비율 측면에서는 국어 문항이 평균 99.73%로 가장 높았으며, 사회 문항과 과학 문항이 각각 99.66%, 98.38%로 그 뒤를 이었다. 이는 사회와 과학 교과에 비해 국어 교과의 분석 대상 문항에 자동채점 처리가 용이한 P1~P2 문항 비중이 높았기 때문인 것으로 추정된다. 또한 분석 대상 문항의 답안 유형 수 측면에서도 자동채점 가능성이 국어, 사회 과학 순으로 나타난 결과를 반영한 것이기도 하다.

그러나 채점 신뢰도 측면에서는 사회 문항의 Kappa계수가 평균 1.00으로 나타나 가장 높은 신뢰도를 보여주었으며 국어 문항과 과학 문항이 .99, .94로 그 뒤를 이었다. 채점 불일치 비율은 채점 신뢰도와 반대로 과학(1.12%), 국어(0.24%), 사회(0.09%) 순으로 나타났는데, 채점자 간 일치도를 보여주는 Kappa계수는 인간채점과 자동채점 간의 불일치 비율과 그 경향을 같이 하기 때문이다.

〈표 9〉 2012년 교과별 채점 결과 평균

교과	정답 패턴	채점 비율	Kappa계수	채점 불일치 비율
국어	P1	99.98	1.00	0.11
	P2	99.86	.99	0.32
	P3	98.07	.97	0.34
	P4	98.17	.98	1.02
	전체	99.73	.99	0.24
사회	P1	100.0	1.00	0.06
	P3	99.82	1.00	0.05
	P4	95.88	.98	0.45
	전체	99.66	1.00	0.09
과학	P1	99.99	1.00	0.11
	P2	99.68	.97	0.67
	P3	99.53	.95	2.10
	P5	88.14	.64	5.05
	전체	98.38	.94	1.12

* 소수점 셋째 자리에서 반올림

교과 간 자동채점 결과의 차이를 보다 구체적으로 살펴보면 다음과 같다. 첫째, 국어 교과의 채점 결과, 채점 비율은 98.1~100%로 매우 높았고, Kappa계수 및 상관계수 또한 최소 .97 이상으로 채점 신뢰도가 매우 높았다. 채점 불일치 비율은 P1, P2 유형에서는 0~0.2%(〈표 6〉 참조)로 매우 낮았으나, 답안의 길이와 복잡도로 인해 3단어 술어형의 P3, P4 유형의 두 문항은 각각 0.34%, 1.02%로 상대적으로 높게 나타났다. 일반적으로 미판단 답안 및 채점 불일치 사례는 정답 외에 다른 단어・구를 부가적으로 삽입하거나, 정답에 해당하는 다양한 유의어나 유사 단어가 존재하는 경우로 나타났는데, 국어 교과에서는 답안의 의미 관계를 살펴 부분 점수를 부여해야 하는 경우에 두드러졌다.

둘째, 사회 교과의 채점 결과, 정답과 유사한 다양한 표현이 존재하는 P3 유형의 세 문항을 제외하고는 채점 비율이 100%에 달하였고, Kappa계수와 상관계수 또한 문항 대부분이 1.00으로 나타났으며 가장 낮은 경우에도 .98 이상으로 채점 신뢰도가 매우 높았다. 채점 불일치 비율의 경우에도 0~0.45% 정도로 낮았다. 이처럼 사회 교과에서 특히 미판단 답안과 채점 불일

치 사례가 적은 까닭은 유의어가 비교적 적으면서도 명확한 개념어를 요구하는 문항이 다수 출제되어 정오답 판정이 비교적 분명하고 채점의 주관성이 개입될 여지가 작기 때문이다. 일반적으로 미판단 답안 및 채점 불일치 사례는 정답 외에 다른 단어·구를 부가적으로 삽입하거나, 정답에 해당하는 다양한 유사 표현이 존재하는 경우로 나타났다. P1 유형이 많은 사회 교과의 경우, 철자 오류와 같은 채점 불일치 사례 중 답안 판단 기준이 단순한 경우 반복 채점으로 인한 채점자 오류가 나타나기도 하였다.

셋째, 과학 교과의 채점 결과, 채점 비율은 88.1~100%로 다른 교과에 비해 상대적으로 낮은 편에 속하고, Kappa계수와 상관계수가 각각 .64~1.00, .81~1.00으로 문항별 편차가 큰 편이었다. 그러나 이는 문장 수준의 프로그램 제고를 염두에 두고 실험적으로 검증한 P5 유형의 문장 수준 답안의 결과에 기인한 것으로, 이를 제외한 과학 문항의 채점 비율은 99.5~100%이고, Kappa계수와 상관계수는 .95 이상으로 채점 신뢰도가 높은 편이었다. 채점 불일치 비율도 P5 유형의 문장 수준 답안 문항을 제외하면 0.10~2.10%로 낮은 편에 속한다. 일반적으로 미판단 답안 및 채점 불일치 사례는 정답 외에 다른 단어·구를 부가적으로 삽입한 경우로 나타났는데, 과학 교과에서는 이 외에도 정답에 대한 유사 답안이 다양하게 존재하는 경우이거나 부분 점수를 부여하는 경우가 두드러졌다.

Ⅴ. 결론

한국어 자연언어처리 기술 및 지식베이스의 여건이 충분하지 않은 상황에서, 비교적 개발이 쉬운 단답형 자동채점 시스템부터 개발하는 것은 큰 의미가 있다. 현재 대규모 평가의 서답형 문항에서 다수를 차지하고 있는 단답형 문항을 처리할 수 있어 그 활용 가치가 높을 뿐만 아니라, 이를 바탕으로 좀 더 장기적으로 내용 기반의 서답형 문항 자동채점도 지속적으로 연구·발전시킬 수 있는 토대를 제공하기 때문이다. 물론 향후 한 문장 단위를 넘어 두 문장 단위 이상의 복문을 채점하기 위해서는 복잡한 구문 분석과 의미 분석이 요구되는데 현재의 한국어 처리 기술과 지식베이스 축적 수준으로 볼 때 이를 완전히 담보하기 어렵다. 기술적 측면에서나 한국어 지식베이스 자원 측면에서 꾸준한 개발 및 보완 노력이 필요하다.

특히 한국어 지식베이스 자원 측면에서 중·장기 개발 방향을 설정하는 데 교과별 채점 도구의 마련에 대해 깊이 고려할 필요가 있다. 앞서 국어, 사회, 과학 교과를 대상으로 살펴본 바와 같이 교과별 사용 용어 및 용례, 문항 출제 형식은 서로 다른 특징을 보이며, 이는 채점 결과에도 영향을 미쳤다. 이를 볼 때, 더 높은 정확성과 채점 비율로 자동채점 프로그램의 성능을 향상시키고자 한다면 교과별 지식베이스(단어, 문장, 사전, 코퍼스, 시소러스, 온톨로지)의 마련

이 절실하다. 정밀한 구문 분석, 의미 분석을 수행할 수 있는 자동채점 프로그램이 개발된다 할지라도 해당 교과 혹은 평가 도구의 용어에 대해 다년간의 지식베이스가 구축되어 있지 않으면 채점의 정확성 및 효율성이 떨어질 수 있다. 반면, 현재의 단어·구 수준 자동채점 프로그램에서도 풍부한 교과별 지식베이스가 구축된다면 채점의 정확성 및 효율성이 크게 향상될 수 있을 것으로 기대된다. 이는 대규모 평가의 서답형 문항들은 대부분 문항의 형식 또는 교육과정에 따른 지문이나 내용을 토대로 학생의 답안 반응을 제약하고 있어 복합적인 분석 기술을 요구하지 않고서도 채점이 가능한 측면이 있기 때문이다. 즉 특정 교과, 지식, 영역을 중심으로 자연언어처리 기술과 지식베이스를 마련해 놓으면 교육과정 상의 내용을 기반으로 하는 학생 답안에 대해서는 정확하면서도 용이하게 자동채점이 가능할 것으로 예상된다.

교과별 지식베이스 자원을 구축하는 것과 별도로 일반적인 언어 정보를 구축해야함은 물론이다. 이는 기존에 언어 정보를 구축하고 있는 관련 연구 기관의 협조가 필수적이므로, 이들 기관과 협동 연구 체제의 구축이 요구된다. 한국어와 관련한 말뭉치나 시소러스를 구축해 온 기관으로 국립국어연구원, 한국과학기술원, 연세대학교, 고려대학교 등이 있는데 이들은 한국어 사전이나 어휘 정보를 다루면서 오랜 기간 말뭉치와 시소러스를 구축하여 방대한 정보를 축적하고 있다. 따라서 관련 기관과의 협조 속에 1차 자료를 수집하여 특정 평가 도구와 교과에 도움을 줄 수 있는 범위로 한정하여 시소러스를 구축하고, 일반 말뭉치를 학습자 언어를 처리할 수 있는 말뭉치로 특화하여 구축해 나갈 수 있다. 이는 본 프로그램을 개발하면서 동시에 얻게 되는 학습자의 답안 자료와 목록, 정답 템플릿 자료 등을 함께 참조하여 구축하면 그 실효성은 더욱 제고될 것이다.

일반적인 언어 정보 및 교과별 지식베이스 구축 외에도, 교과별 언어처리 기술 및 개념 분석 기술이 개발될 필요가 있다. 향후 단어·구 수준이 아닌 문장 이상의 답안을 유효하게 처리하기 위해서는 복잡한 구문 분석과 의미 분석, 클러스터링, 유의어 데이터베이스의 확장 적용 등이 요구된다. 구체적으로, 답안의 품사 형태 및 후속 구절과의 호응 등 구문 및 의미 분석이 세밀하게 이루어져 그에 따라 부분 점수가 부여되는 채점 과정이 포섭될 수 있도록, 향후 문장 수준의 자동채점 프로그램에서는 개념 기반 채점의 단계가 보다 정교화될 필요가 있다. 또한 문항에서 언어 단위의 제한이 없으며 동사의 활용형이 다양한 답안에 대해서 의미 관계까지 포섭할 수 있는 클러스터링 방안을 정교화할 필요가 있다. 마지막으로 답안 내용 가운데 개념으로 설정된 핵심 용어 외의 답안 내용을 검토할 수 있도록 하는 옵션을 설정하거나 유의어 데이터베이스를 확장·적용하는 등의 방안을 통해 자동채점 프로그램을 보완할 필요가 있다.

그런데 구문 분석과 의미 분석, 클러스터링, 유의어 데이터베이스의 확장 적용 등 언어처리 및 개념 분석 기술을 개발하고자 할 때, 교과별 지식베이스와 연계되어 차별화된 언어처리 및 개념 분석 기술이 개발되는 것이 유용하다. 각 교과마다 특별히 자주 사용되는 관용적 표현이 존재한다. 가령, 과학 교과의 '끓는 점이 100℃이하이다.', 사회 교과의 '수요곡선이 우하향한다.'라는 표현은 해당 교과의 교수·학습 내용에서 자주 언급되는 표현으로, '수요곡선'이라는 주

어와 '우하향한다'라는 동사는 자주 결합되어 나타난다. 일상 언어적 표현이 많은 국어 교과는 보다 광범위하고 정밀한 구문 분석과 의미 분석이 요구되는 반면, 이들 교과는 언어 처리 및 개념 분석 기술을 개발할 때 교육과정이나 교과서에 자주 등장하는 관용적 표현과 그 변용적 표현만 고려해도 채점의 정확성 및 효율성을 상당한 정도로 높일 수 있을 것이다.

지금까지 자동채점 프로그램의 향후 발전 방향성을 검토하기 위해 교과 간 문항의 특성 및 자동채점 결과의 차이를 살펴보았다. 이러한 교과 비교가 향후 한국어 서답형 자동채점 프로그램을 개발함에 있어서 교과별 채점 정확도를 제고하는 데 기여할 수 있을 것으로 기대한다. 본 연구를 바탕으로 대규모 평가의 서답형 문항 채점에 소요되는 시간적·경제적·행정적 부담이 자동채점을 통해 얼마나 절감되는지 비교하여 자동채점 프로그램의 효율성을 검증하는 후속 연구가 요청된다.

참 고 문 헌

강원석(2011). 질의문 유형 분석을 통한 서답형 자동채점 시스템. **한국콘텐츠학회논문지, 11**(2), 13-21.

교육과학기술부(2010. 5). **창의성과 인성 함양을 위한 교육내용・방법・평가체제 혁신 방안 VIP 보고**. 대통령 주재 제3차 교육개혁 대책회의(2010년 5월 19일).

교육과학기술부(2011. 3). 2011 창의・인성 교육 기본 계획. 보도자료(2011년 3월 11일).

권오영(2004). 웹 기반 주관식 평가문항 채점 알고리즘 설계 및 구현. 한서대학교 교육대학원 석사학위 논문.

김경성, 김종훈, 곽현석(2011). **2012년 국가수준 학업성취도 평가 채점 표준화 방안 연구**. 한국교육과정평가원 교육과학기술부 수탁과제 2011-1.

김경희, 김완수, 김동영, 김종훈, 김미경, 최인봉, 신동광, 박인용, 이인호, 신진아, 최인선, 송미영, 한정아, 김희경, 한경택, 박거도(2013). 컴퓨터 기반 국가수준 학업성취도 평가 도입 방안. 한국교육과정평가원 연구보고 CRE 2013-5.

김미경, 김도남, 김영란, 김현정, 이정우, 서민철, 조윤동, 조성민, 최인선, 김동영, 이인호, 이영주, 고현숙(2012). 2012년 국가수준 학업성취도 평가 출제 연구. 한국교육과정평가원 연구보고 RRE 2012-2-1.

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

노은희, 김명화, 성경희, 김학수(2013). 대규모 평가를 위한 서답형 문항 자동채점 프로그램 정교화 및 시범 적용. 한국교육과정평가원 연구보고 RRE 2013-5.

박희정, 강원석(2003). 유의어 사전을 이용한 주관식 문제 채점 시스템 설계 및 구현. **한국컴퓨터교육학회논문지 6**(3), 207-216.

성태제(2014). **현대교육평가**(제4판). 서울: 학지사.

성태제, 양길석, 강태훈, 정은영(2010). 학업성취도 평가 서답형 문항 컴퓨터 채점화 방안 탐색. 한국교육과정평가원 연구보고 RRE 2010-1.

성태제, 이양락, 시기자, 이경언, 이근호, 박태준, 노원경, 박찬호, 박도영, 정은주(2013). 행복교육, 창의인재 양성을 위한 교육과정, 교수・학습, 교육평가 패러다임 전환 (pp. 26~50). 성태제 외(공저), **2020 한국 초・중등교육의 향방과 과제 -교육과정, 교수・학습, 교육평가-**. 서울: 학지사.

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

교육과정평가원 연구보고 RRE 2012-10.

신동광, 민호기, 박상복, 정채관, 주헌우, 김미지, 김연희, 이은숙, 김동남, 김영준(2012). 국가영어능력평가시험의 말하기 자동채점 프로그램 도입 방안. 한국교육과정평가원 연구보고 RRE 2012-9.

정동경(2001). 벡터 유사도와 시소러스를 이용한 주관식 답안의 채점 방법. 동국대학교 교육대학원 석사학위 논문.

조우진(2006). 의미 커널과 한글 워드넷에 기반한 지능형 채점 시스템. 한림대학교 대학원 석사학위 논문.

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.

Attali, Y. (2011). A Differential Word Use Measure for Content Analysis in Automated Essay Scoring. ETS Research Report ETS RR-11-36.

Ben-Simon, A., & Bennett, R. E. (2007). Toward More Substantively Meaningful Automated Essay Scoring. *Journal of Technology, Learning, and Assessment*, *6*(1).

Butcher, P. G., & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers & Education, 55*, 489-499.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment, 5*(1), 4-35.

Landis, J. R., & Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Meadows, M., & Billington, L. (2005). *A Review of the Literature on Marking Reliability*. National Assessment Agency.

NAEP(2010). What to Expect for the NAEP Writing Computer-Based Assessment. NCES 2010-470.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*(1), 183-196.

Shermis, M. D., & Burstein, J. (2003). *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates, Inc., Publishers. Mahwah, New Jersey.

• *논문접수 : 2014-04-30/ 수정본접수 : 2014-05-29/ 게재승인 : 2014-06-13*

ABSTRACT

A Comparative Analysis of Scoring Results in Korean Automatic Scoring Program for Short-answer Items - focused on the three subjects in NAEA: Korean, Social Studies and Science -

Eun-Hee Noh, Kyung-Hee Sung
(Research Fellow, Korea Institute for Curriculum and Evaluation)

The purpose of this study is to analyze the differences of scoring results and answer types among subjects in the 2012 NAEA(National Assessments of Educational Achievement) using the KASS(Korean Automatic Scoring System) developed in 2013. The subjects are Korean(17 items), Social Studies(13 items), Science(8 items) and the numbers of answer are 3,010 of each subject. First, in supply-type items of 2009~2012 NAEA, the rates of answer types are composed of short-answer 74.7%, a sentence 12%, multi-sentence 10.9%, and the others 2.4%. Considering each subject, the social studies showed the highest rate(86.5%) in case of short-answer(P1~P3) while the Korean revealed the highest rate(17%) in case of sentence-level answers(P4~P6) compared to the other subjects. That is to say, questions asking predicate words, phrase and sentence-level answers have mostly been on the Korean test. However, questions asking content-focused concept words have been on the social studies test. Second, the result of scoring indicate that Kappa coefficients of short-answer items were high above .95, but the longer and more complicated length of answers was, the less correlation coefficient between human scorer and KASS was. Moreover, the Korean showed the highest rate(99.73% on average) in terms of scoring rates, but the social studies showed the highest value(Kappa coefficient 1.00 on average) in terms of scoring reliability.

To sum up, terminology and its usage as well as questions forms of each subject have different features, which finally affects the scoring results. Therefore, if knowledge-based system according to each subject was constructed and differentiated natural language processing technology was sophisticated, accuracy and efficiency of the automatic scoring program could considerably improve.

Key Words : automatic scoring, Korean automatic scoring program, supply-type items, National Assessments of Educational Achievement