국가수준 학업성취도 평가 국어 서답형 문항의 자동채점 결과 분석

노은희 한국교육과정평가원

* 이 연구는 '대규모 평가를 위한 서답형 문항 자동채점 프로그램 정교화 및 시범 적용' 한국교육과정평가원 연구보고(RRE 2013-5)의 일부로서, 국어교육학회 제55회 전국학술대회에서 발표한 내용을 재구성한 것임을 밝힌다.

I. 서론

최근 교육계 전반에서 선택형 평가 방식을 벗어나기 위한 노력이 분주하다. 2009 개정 교육과정 총론에서는 학생들의 창의성과 인성 함양을 표방하며, 평가 방법 측면에서도 '교과의 평가는 선택형 평가보다는, 서술형이나 논술형 평가 그리고 수행 평가의 비중을 늘려서 교과별 특성에 적합한 평가를 실시하도록' 권고하고 있다. 교육과학기술부는 이에 발맞추어 '창의성과 인성 함양을 위한 교육내용 · 방법 · 평가체제 혁신방안 VIP 보고'(2010년 5월 19일), '2011 창의 · 인성 교육 기본 계획'(2011년 3월 11일), '중등학교 학사 관리 선진화 방안'(2011년 12월 14일) 등의 발표를 통해, 선택형 평가에서 탈피하여 서술형 평가로의 전환을 강력히 요구하며 단계적으로 2013년까지 시도별 · 교과별 특성을 고려하여 일정 수준(20~40%) 이상으로 서술형 평가를 확대하도록 권장하고 있다. 이러한 일련의 흐름은 선택형 평가가 효율적인 인력 배치가 우선적으로 필요한 시기에는 유용한 방식이나 향후 국가 경쟁력을 제고할 창조적 인재상을 키우기에는 한계가 있다는 지적에 따라, 이제 교육 전반의 평가 방식을 전환해야 한다는 인식에 터한다.

그런데 선택형 평가에 대비되는 '서술형' 평가라는 용어는 흔히 교육과정 문서는 물론 교실 현장에서 통상 쓰고는 있으나 명확하게 규정된 바는 없다. 즉, '서술형'이란 용어는 교육과정 및 교육 현장에서 '문장 이상으로 학생이 응답하는 문항 형태'에 대해 관례적으로 사용하고 있으나, 교육평가 분야의 학문적 용어로 제대로 정립된 것은 아니다. 문항 유형으로 정립된 학문적 용어는 '서답형'으로, 굳이 따지자면 서술형은 이러한 서답형 문항의 하위 유형에 해당한다고 할 수 있다. 서답형 문항(supply item)은 피험자가 주어진 답지에 따라 정답을 선택하는 것이 아니고 직접 정답을 구성하여 작성하는 문항 형태이다(한국교육평가학회, 2004: 207). 보통 서답형은 단답형, 괄호형, 논술형(또는 서술형) 등의 세부 유형으로 구분지어 볼 수 있다. 여기서 '논술형'은 학생이 나름대로의 생각이나 주장을 논리적으로 설득력 있게 조직하여 작성해야 함을 강조하는 형식이며, '서술형'은 논술형에 비해 비교적 짧으면서 객관적인 정답이 존재하는 형식 정도로 구분하여 이해하기도 한다. 따라서 본고에서는 한 문장 이내의 단답형인 서답형을 주 대상으로 하므로 서술형이란 통상적인 용어보다는 서답형이란 용어를 사용하고자 한다.

서답형 문항은 수험자가 답안을 직접 작성하기 때문에, 반응의 자유가 있어서 선택형 문항에 비해 고등정신 능력을 측정하는 데 효과적이다. 반면에 평가 기준 작성의 어려움, 채점 시간, 채점 피로도, 채점 결과의 공정성 문제로 지금까지 교실 현장에서 교사가 출제를 기피하는 측면이 있었던 것이 사실이다. 특히 대규모 평가에서는 과다한 채점 비용 발생과 채점자들 간의 변인 및 신뢰도 차이로 서답형 문항을 활용하는 데는 여러 제약이 따른다.

현재 국가수준 학업성취도 평가(이하 학업성취도 평가)도 서답형 문항 채점의 어려움으로, 간단한 단답형에서부터 논술형까지의 서답형 문항 가운데에서 매우 높은 비율로 단어 · 구 수준의 단답형 문항을 활용하고 있다.[1]

---

1 우리와 비교하여 국외 대규모 평가에서의 서답형 문항 출제 비율에 대해 살펴보면, 국제 학업성취도 평가인 PISA와 TIMSS, 자국내 국가수준 학업성취도 평가인 영국의 NCA, 미

학업성취도 평가에서 전수평가가 안정적으로 정착된 2009~2012년 최근 4년 동안의 초6, 중3, 고2 국어 서답형 문항 현황을 살펴보면 〈표 1〉과 같다. 여기서 단어 · 구 답안은 전체 서답형 문항 중 66%이고, 한 문장 답안과 다문장 답안은 각각 17%로 나타난다.[2]

표 1. 국어 서답형 문항의 답안 길이별 문항 수 및 비율(2009~2012년)

전체 문항 수*	단어 · 구 답안		문장 답안		다문장 답안	
	문항 수	비율	문항 수	비율	문항 수	비율
165(100%)	109	66%	28	17%	28	17%

* 문항 수는 하위 문항 수를 기준으로 산정함.

국어 교육 목표에 비추어 고등 정신 기능과 국어 능력의 질적인 측면을 평가하려면 사실 단어 · 구 수준의 답안보다는 한 문장 이상의 답안 형태가 교육의 본질적 측면에 도달하는 데 좀 더 유익할 것이라는 예측은 가능하다. 그럼에도 불구하고 대규모 평가 시 서답형 문항은 채점의 어려움으로 문장 이상의 서답형 문항을 출제하기 어려운 현실적 한계에 봉착하게 된다. 대규모 평가에서 활용할 수 있는 서답형 채점 방안은 크게 온라인 인간채점 방식과 컴퓨터 자동채점 방식을 생각해 볼 수 있다. 현재 학업성취도 평가의 채점 방식인 온라인 채점의 경우, 2012년에 50만 고2 응시자의 국어, 수학, 영어 답안 채점과 관련하여 7월 19일부터 8월 14일까지 한 달여 동안 채점자 3,689명, 관리자 115명이 동원되어 총 비용 14억 원가량을 소요하였다(김미경 외, 2012: 116-121 참조). 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,

---

국의 NAEP, 호주의 NAPLAN 등은 전체 문항 수 중 서답형 비율이 30~60%로, 현재 한국 학업성취도 평가의 20% 정도에 비해 그 비율이 훨씬 높다(노은희 외, 2012: 16-18).

2 다른 교과의 경우를 보면, 단어 · 구 답안이 사회는 86.5%, 과학은 77.7%로 국어보다 훨씬 높은 비중을 차지한다.

2011: 8-22 참조). 즉, 한 해 초 · 중 · 고 학업성취도 평가의 서답형 채점으로 대략 총 63억 원의 예산과 약 11,000명 이상의 인원이 동원된 것이다.[3] 이를 고려할 때, 대규모 평가에서 서답형 문항의 실제적 활용을 도모하기 위해서는 채점에 소요되는 시간적 · 경제적 · 행정적 부담을 최소화할 수 있는 방안을 강구할 필요가 있다.

한편, 오늘날 세계 각국은 컴퓨터 및 인터넷을 활용하는 평가 체제의 도입을 적극적으로 모색하고 있다.[4] 이러한 국제적인 평가 동향에 발맞추어 우리나라에서도 학업성취도 평가에서 컴퓨터 기반 평가 시스템 도입을 고려하고 있다.[5] 향후 대규모 평가들이 컴퓨터 기반으로 시행된다면 서답형 문항의 자동채점 프로그램은 필연적으로 요구되는 시스템이다. 현재의 학업성취도 평가와 같이 교과 내용 기반으로 답안이 제한적이고 그 언어 단위 길이가 짧을수록 기계에 의한 자동채점 가능성과 채점 신뢰도는 높아지며, 특히 대규모일수록 예산 절감 효과가 커진다. 따라서 우선 현재의 언어 처리 기술로 해결 가능한 단어 · 구 수준의 짧은 답안부터라도 자동으로 채점할 수 있는 프로그램을 개발하고 이를 적용하는 연구를 시도할 필요가 있다. 이에 노은

---

3 2012년 학업성취도 평가의 서답형 채점은 초6~중3 답안의 경우 시도교육청에서, 고2 답안의 경우 한국교육과정평가원에서 담당하였다.

4 미국에서 추진하고 있는 차세대 학력평가 2.0의 주요 프로그램 SBAC와 PARCC는 CAT(Computer Adaptive Test) 혹은 CBT(Computer Based Test)로 학생들의 학업 능력 및 향상도를 평가한다(성태제 외, 2013: 43 참조). 또한 국제 수준의 대규모 학업성취도 평가인 PISA와 ICILS에서도 점차 컴퓨터 기반 평가를 도입하고 있는데, PISA의 경우 2006 과학 소양, 2009 디지털 읽기 소양, 2012 문제해결능력 평가 등에서 컴퓨터 기반의 평가 방식을 통해 학생들의 기본 소양을 측정해 왔으며 PISA 2015부터는 모든 영역에서 컴퓨터 기반 평가를 도입하는 것을 추진하고 있고, 2013년에 처음 시행되는 ICILS에서는 실제 웹기반 환경과 동일한 상황에서 정보를 검색 · 선택 · 각색하는 능력을 평가한다(송미영 외, 2013: 9-10 참조).

5 교과부는 '스마트교육 추진 전략(2011.6.)'에 따라 온라인 평가 체제를 구축하여 학업성취도 평가를 IBT 방식으로 전환하는 계획을 발표하였다. 이에 김경희 외(2013)에서는 '컴퓨터 기반 국가수준 학업성취도 평가 도입 방안'을 연구하였다.

희 외(2012; 2013)는 2012년부터 대규모 평가에서 서답형 문항의 활용을 제고하고 채점의 효율화를 도모하기 위해, 단어 · 구 수준의 한국어 서답형 문항 자동채점 프로그램을 개발 · 적용하는 연구를 진행하였다.

본 연구는 2013년 개발된 한국어 서답형 문항 자동채점 프로그램 KASS(Korean Automatic Scoring System)을 활용하여 2012년 학업성취도 평가의 초 · 중 · 고 국어 총 17문항 각 3,010개 답안을 대상으로 자동채점 결과를 분석하여 향후 국어 서답형 문항의 자동채점 가능성을 모색해 보고자 한다.

II. 한국어 서답형 문항 자동채점 프로그램 개관

1. 자동채점 프로그램 대상 문항 분석

영어권에서는 대규모 평가의 서답형 문항에 대해 기계로 자동채점하는 연구 및 적용이 활발하다. 이에 비해 한국어의 경우 이와 관련한 연구가 매우 미진한 편이다.[6] 이러한 상황에서 노은희 외(2012, 2013) 연구에서 개발한 한국어 서답형 문항 자동채점 프로그램은 아직 단어 · 구 수준의 짧은 답안을 대상으로 하고는 있으나, 대규모 평가에서 한국어의 서답형 답안을 기계로 채점하는 시스템을 최초로 구축하였다는 점에서 일차적인 의의를 지닌다.

현재 한국어 처리 기술이 아직은 충분하지 않고 지식베이스가 필요한 만큼 축적되어 있지 않기 때문에, 한국어 처리 기술의 진척 정도에 따라 채점 가능한 서답형 답안을 미리 판정하는 작업이 필요하다. 이에 한국어 서답형 답안을 대상으로 문자열 일치, 형태소 분석, 구문 분석, 의미/담화 분석 등

6 한국어 자동채점 프로그램 개발과 관련하여 특정 영역의 소규모 시험을 대상으로 실험실 수준의 몇몇 시도는 있었으나(정동경, 2001; 박희정 · 강원석, 2003; 권오영, 2004; 조우진, 2006; 강원석, 2011 등), 실험 결과가 타당도와 신뢰도 측면에서 검증되지 않았고 대규모 평가나 다른 영역에서 활용 가능한지는 추후 논의가 필요한 상태이다.

의 자연언어 처리 기술 필요 여부에 따라 정답 패턴을 P1에서 P6까지 구분하여 보고자 한다. 〈표 2〉는 이를 간략히 정리한 것이다.

표 2. 한국어 서답형 문항의 정답 패턴 분석과 예시

구분	P1	P2	P3	P4	P5	P6
언어 처리 기술*	문자열 일치	형태소 분석 (어휘 토큰 분석)	형태소 분석 (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
용어 수**	1~2	2~3	3~4	4~6	6~8	8~10
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.

* P2부터 이전 정답 패턴에 사용된 언어 처리 기술이 누적되어 사용됨.

** 한국어 조사, 어미를 제외하고 띄어쓰기를 하는 주요 내용어의 수를 의미함.

P1은 형태소 분석 없이 정답과 문자열이 완전하게 일치하면 자동채점이 가능한 유형이다. P2는 형태소 분석 중 어휘 토큰[7] 분석이 필요한 것으로, 부분 문자열만 일치하면 자동채점하는 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. P4부터는 문장 수준 이상의 답안을 대상으로 개념(concept)[8]이 동일한지를 판단하는 구분 분석이 필요하다.

---

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

8 '개념'은 컴퓨터에서 특정 문항의 채점 정보를 저장할 때 특정 주제를 기술하는 단어(들)

P4는 단순 개념의 구분 분석이, P5는 복합 개념의 구분 분석이 요구된다. P6의 경우에는 복문 및 다문장 수준 답안을 채점하기 위해 의미와 담화 분석이 요구되는 경우이다.

2. 자동채점 프로그램 시스템 구조

2013년 개발된 자동채점 프로그램은 단어 · 구 수준 P1~P3의 답안 처리를 목표로 한다. 〈그림 1〉은 2012년의 시스템 설계를 기초로 2013년에 정교화하여 개발한 자동채점 시스템 구조도이다. 자동채점 시스템은 크게 '답안 분석 및 정규화' 과정, '자동채점' 과정, '후처리 수작업 채점'의 3과정으로 구성된다.

먼저, '답안 분석 및 정규화' 과정에서는 채점 대상이 되는 학생 답안이 입력되면 채점 옵션에 따라 답안을 분석하여 정규화를 수행한다. 정규화(normalization)는 다듬어지지 않은 데이터를 컴퓨터가 효율적으로 처리할 수 있도록 일정한 규칙에 따라 변형하는 것을 의미한다. 본 프로그램에서는 채점 기준에서 띄어쓰기가 의미 없는 경우의 띄어쓰기 교정, 정/오답 판단에 영향을 미치지 않는 공백 및 기호 제거 등이 옵션으로 설정되어 있어 문항에 따라 이를 선택적으로 활용하여 정규화를 수행할 수 있다.

다음으로 '자동채점' 과정에서는 정답 템플릿[9]에 기술된 채점 기준에 따라 채점을 수행하고 결과를 저장한다. 구체적으로 자동채점 과정은 모범 답안 일치 채점, 고빈도 답안 일치 채점, 개념 기반 채점, 단서어 기반 오답 처리의 순서로 진행되며 각 과정에서 모든 답안을 처리하고 미채점된 답안들만 다음 단계로 넘어간다. 모범 답안과 고빈도 답안 일치 채점은 정답 템플

정도로 이해할 수 있다. 하나의 개념은 하나의 단어, 구, 문장에 대하여 기술이 가능하도록 n개의 Word라는 단위를 갖는다.

9 정답 템플릿은 특정 문항의 채점 기준과 정보를 저장하고 있는 파일이다.

릿의 해당 정보와 학생 답안이 완전하게 일치하는 경우에 점수를 부여한다. 개념 기반 채점은 학생 답안의 일부 혹은 전체가 정답으로 인정할 수 있는 개념과의 일치 및 유사성 여부 정도를 판단하여 점수를 부여하는 단계이다. 개념 기반 채점으로 걸러지지 않는 답안은 향후 수작업 채점으로 넘어가는데, 이때 '단서어 목록'을 작성하고 답안에 단서어(cue word)[10]가 출현하지 않으면 오답으로 처리하는 과정을 추가한다. 단서어 기반 오답 처리는 정답 템플릿에 단서어가 존재하는 경우에만 수행되며 학생 답안에 단서어 목록의 단어가 하나도 존재하지 않으면 오답으로 처리한다.

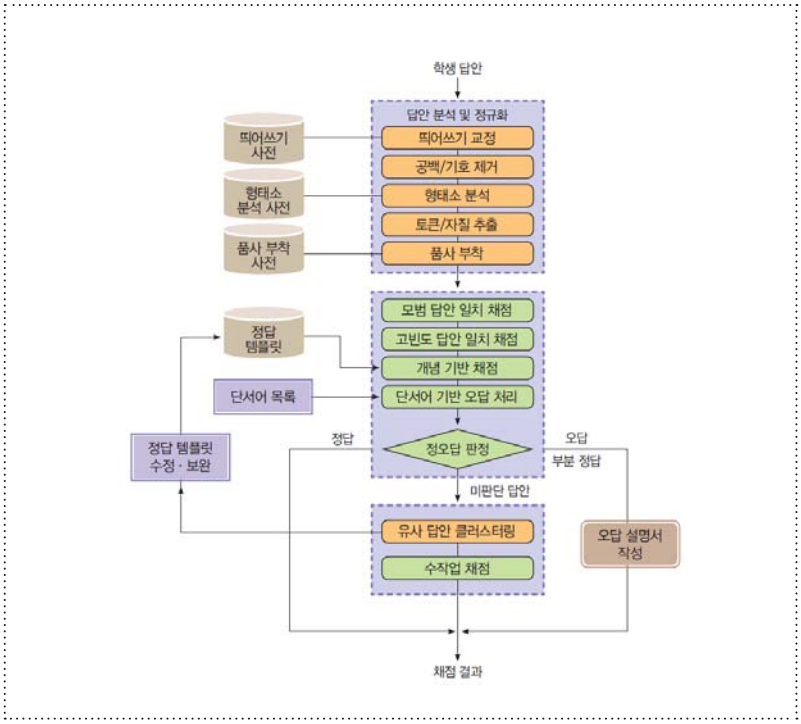

그림 1. 2013년 개발 한국어 서답형 문항 자동채점 시스템 구조도

10 단서어(cue word)는 정답 또는 부분 정답이 될 수 있는 단어로 이것이 출현하기만 하면 무조건 오답으로 처리하지 않고 수작업 채점을 할 수 있도록 미판단 답안으로 남겨 두는 것이다. 이는 자동채점의 오류 가능성을 줄이기 위한 단계라 할 수 있다.

자동채점이 끝난 후 채점되지 않은 미판단 답안들은 '후처리 과정'으로 넘겨진다. 후처리 과정에서는 비슷한 답안들끼리 모아주는 클러스터링[11] 작업을 수행하여 정답 템플릿을 갱신하거나 빈도가 낮은 답안 유형들을 수작업으로 채점한다.

사실, 자동채점이 갖는 용어의 유인으로 학생 답안만 입력하여 넣으면 저절로 답안이 채점되어 나온다는 인식을 갖기 쉽다. 그러나 기계가 채점하려면 채점 기준이 되는 정답 템플릿을 사람이 입력하여 주고 자동채점 단계에서 거르지 못한 미판단 답안은 교과 전문가가 처리해야 한다. Dube & Ma(2010: 16)에 따르면, 자동채점 프로그램의 시스템은 다양하지만 대개 채점자가 ①채점 기준표 제공, ②샘플 답안을 이용하여 모범 답안 템플릿 생성, ③샘플 답안과 비교하여 학생 답안 분석 및 채점, ④채점 과정을 확인하기 위한 수작업 조정 등에서 사람 손을 거친다. 이러한 절차 속에서 채점 관리자가 개입하여 정답 템플릿을 생성하거나 조정하는 것이며, 프로그램은 이를 보다 간단하고 효율적으로 진행할 수 있도록 도구들을 제공한다. 즉, 기본적으로 자동채점 프로그램이라고 해서 컴퓨터가 완전하게 100% 처리하는 것이 아니라 채점 관리자가 주요 단계마다 개입하여 채점 과정을 진행하는 것이고, 이러한 인간 개입 과정을 최소화하는 과정이 곧 자동화 과정인 것이다.

〈그림 2〉는 프로그램의 실행을 돕는 도구 중에서 자동채점 도구의 메인 실행 화면을 보여 준다. 교과 전문가는 이러한 사용자 인터페이스 상에서 학생 답안을 불러들여 자동채점을 수행하고 이후 미판단 답안을 처리할 수 있다.

11 클러스터링은 철자, 발음, 의미 등의 유사도에 기초하여 개별 데이터를 그룹으로 묶는 기법을 말한다.

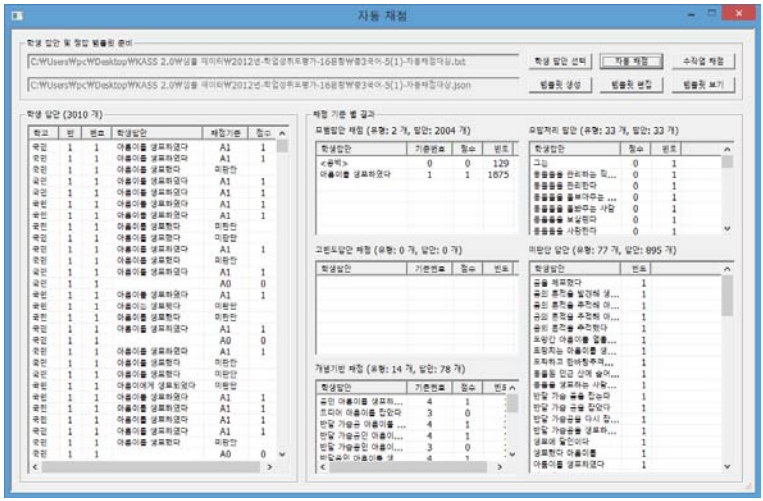

그림 2. 자동채점 프로그램 사용자 인터페이스

III. 학업성취도 평가 국어 서답형 문항 자동채점 결과 분석

1. 자동채점 대상 문항

향후 대규모 평가에서 자동채점 프로그램을 정교화하고 외연적 확장을 꾀하고자 할 때는 교과별 특수성을 고려할 필요가 있다. 즉, 해당 교과가 기반하고 있는 학문 분야에 따라 주로 사용하는 지식베이스와 학습자 용어가 다르기 때문이다. 자동채점 프로그램은 내용 기반으로 정답이 제한적일 때 그 채점 비율이 높아지는데, 국어 교과는 하위 영역 가운데 지식 부분에 해당하는 내용 범주는 자동채점하기 용이하나 그 외 내용 범주에서는 용어나 개념을 딱히 한정하기 어려워 다른 교과에 비해 채점이 까다롭다고 할 수 있

다. 국어 교과의 서답형 문항은 대체로 여러 분야의 말과 글을 이해한 다음, 명사형 단어 외에도 술어형 단어나 구, 문장 형태의 답안을 요구하는 문항이 다수 출제되는데 이것이 교과 고유의 내용이나 용어로 한정되기보다는 다양한 지문과 연결 지어 관련 내용에 학생이 답안을 작성하는 방식으로 문항이 출제되기 때문이다. 따라서 국어 문항에 대한 시범 적용 결과는 자동채점 프로그램의 대상 교과를 확장하는 데 도전적인 의미를 지닌다.

자동채점 프로그램을 검증하기 위해, 2012년 학업성취도 평가의 초 · 중 · 고 국어 총 17문항(하위 문항 기준) 각 3,010개 답안을 대상으로,[12] 자동채점의 채점 비율을 살펴보고, 표집 채점을 통해 확정된 기준 점수와 자동채점 점수 간 일치도, 채점 불일치율을 분석하고자 한다. 현재 학업성취도 평가는 지필평가 방식으로 시행되므로 학생들이 수기로 작성한 답안을 별도로 코딩한 후, 이를 자동채점 프로그램으로 채점한다.

〈표 3〉은 자동채점을 적용할 국어 문항에 대한 정보를 정리한 것이다. 정답 패턴별로 보면 P1 유형은 10문항, P2 유형은 5문항, P3 유형은 1문항, P4 유형은 1문항이다. 이 가운데 P1~P3 유형은 단어 · 구 답안에 해당하며 이는 형태소 분석 처리 기술로도 채점 가능하다. 그런데 P4 유형의 1문항의 경우 문장 수준 답안으로 이를 자동채점 절차로 온전히 채점하려면 구문 분석 처리 기술까지 필요하나, 형태소 분석 처리 기술로 어느 정도까지 채점이 가능한가를 살피기 위해 실험적으로 대상 문항으로 포함한다. 이는 향후 프로그램 개발을 위한 점검 차원이기도 하다.

자동채점의 가능성은 답안 작성 유형,[13] 답안 유형 수[14](1,000개당), 정답

12 2012년 학업성취도 평가에서 2단계 비례층화 군집 표집하여 채점한 약 10,000개의 학생 답안 중, 국어 17문항을 대상으로 각 3,010명의 답안을 학교별 군집 표집하여 추출한다.

13 답안 작성 유형은 모범 정답의 언어 단위 길이를 중심으로 제시한 것이다.

14 답안 유형 수는 정/오답을 포함한 학생 답안의 모든 가능한 유형 수인데, 여기서는 답안 정규화 과정을 통해, 채점 대상이 되지 않는 특수기호(문장부호 등) 제거, 띄어쓰기 무시, 불용어(不用語; '이모티콘'이나 '모름' 등과 같이 답안 판정과 무관한 단어) 제거 등을 통해 사전 정련하여 실제 학생 답안의 다양성보다 그 수가 훨씬 줄어든 것이다.

표 3. 2012년 국어 자동채점 대상 서답형 문항 정보

학교급	번호	답안 작성 유형	답안 유형 수 /1,000	정답 패턴	모범 정답
초6 (5문항)	2-ⓐ	1단어 명사형	10	P1	송편
	2-ⓑ	1단어 명사형	10	P1	곡식
	2-ⓒ	1단어 명사형	15	P1	무
	5-ⓐ	1단어 술어형	55	P2	밝게(환하게, 환히)
	5-ⓑ	1단어 술어형	46	P2	크게(커)
중3 (4문항)	1-ⓐ	1단어 술어형	107	P2	타당(적절)
	1-ⓑ	1단어 명사형	62	P2	근거
	5-(1)	2단어 술어형	7	P3	아름이를 생포하였다.
	5-(2)	1단어 명사형	53	P1	아름이
고2 (8문항)	1-ⓐ	1단어 명사형	6	P1	대조
	1-ⓑ	1단어 명사형	8	P1	강조
	4-(1)-ⓐ	1단어 명사형	12	P1	㉠
	4-(1)-ⓑ	1단어 술어형	69	P2	알맞은
	4-(2)	3단어 술어형	311	P4	사회 구조가 복잡하고
	5-㉠	1단어 술어형	30	P1	높이
	5-㉡	1단어 술어형	33	P1	깊이
	5-㉢	1단어 술어형	90	P1	평평한

패턴을 복합적으로 고려하여 결정한다. 이 가운데 답안 유형 수를 중심으로 자동채점 가능성을 살필 때, 현재의 자동채점 시스템을 고려하여 답안이 단어 · 구이면서 답안 유형 수가 1,000개당 100개 미만인 경우 자동채점 가능성이 '매우 높음', 100~200개인 경우 자동채점 가능성이 '높음', 200~400개인 경우 '보통', 400~600개인 경우 '낮음', 600개 이상인 경우 '매우 낮음'으로 분류한다. 예컨대, 〈표 3〉의 문항 가운데, 1단어 명사형이면서 답안 유형 수가 6개이고 정답 패턴이 P1이어서 문자열 일치로 처리 가능한 고2-1-ⓐ 문항은 자동채점 가능성이 '매우 높음'이고, 3단어 술어형이면서 답안 유형 수가 311개이고 정답 패턴이 P4인 경우인 고2-4-(2) 문항은 자동채점 가

능성이 '보통'이라고 예측할 수 있다. 대체로 1단어라도 명사형인 경우보다 술어형인 경우에 답안 유형 수가 많아진다. 다만, 이러한 기준은 절대적인 것이 아니라 향후 자동채점 프로그램의 정확도가 향상되면 언제든 조정 가능한 것이다.

2. 자동채점 결과

대표 사례로, 2012년 학업성취도 평가 국어 문항 중 자동채점의 양상이 다른 두 문항에 대해 채점 단계별로 채점 비율을 제시하면 다음과 같다. 먼저, 고2-5-㉢의 문항에 대해 자동채점 단계별 비율 및 비율 변화를 순서대로 제시한다.

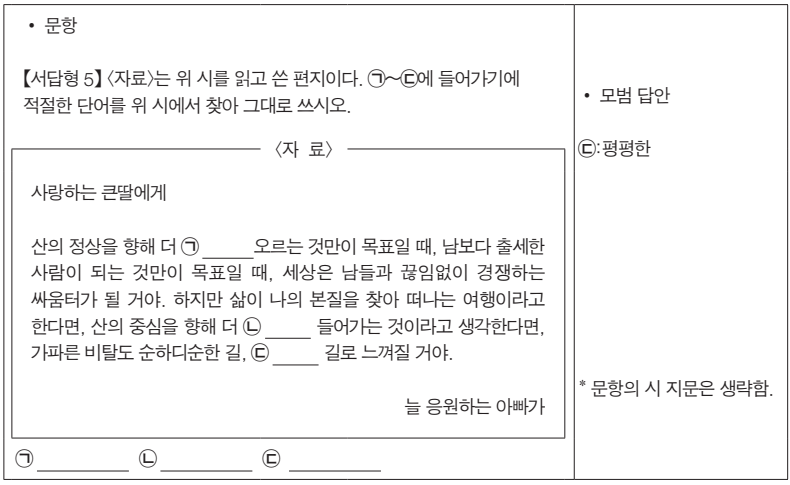

표 4. 고2-5-㉢ 문항의 자동채점 단계별 채점 비율

채점 단계	정답 수	오답 수	채점 수	누적채점 수	미판단 수	채점 비율
모범 답안 일치 채점	1943	320	2263	2263	747	75.2%

고빈도 답안 일치 채점	167	452	619	2882	128	95.7%
개념 기반 채점	1	0	1	2883	127	95.8%
단서어 기반 오답 처리	0	127	127	**3010**	**0**	**100%**
미판단 답안 수작업 채점			0	3010	0	100%

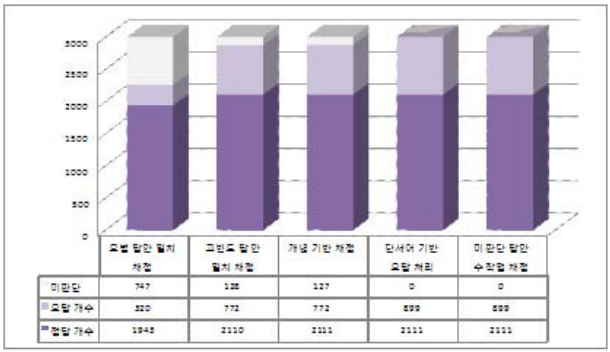

그림 3. 고2-5-ⓔ 문항의 자동채점 단계별 채점 비율 변화

고2-5-ⓔ 문항은 시 지문에서 관련 1단어 술어형 답안을 찾아 쓰는 것으로 응답 자유도가 낮은 편이어서 1,000개당 서로 다른 학생의 답안 유형 수가 90개 정도이다. 정답 패턴도 P1 유형이어서 자동채점 가능성은 매우 높은 문항으로 예상할 수 있다.

실제로 자동채점을 수행한 결과, 채점 단계별 채점 비율을 살펴보면 모범 답안 일치 채점 단계(정답 수 1,943, 오답 수 320)에서 약 75%가 채점되고, 고빈도 답안 일치 채점 단계(정답 수 167, 오답 수 452)에서 약 20%가 채점되어, 약 95%의 대부분 답안이 이 두 단계에서 채점되었다. 개념 기반 채점 단계(정답 수 1, 오답 수 0)와 단서어 기반 오답 처리 단계(정답 수 0, 오답

수 127)에서 약 5%의 채점 비율이 증가하여 모든 답안이 수작업 채점 단계로 넘어가지 않고 채점 완료되었다.

채점자와 자동채점 간 결과가 불일치한 답안 수는 6개(약 0.2%)로, 채점자가 1점을 주었는데 자동채점 프로그램은 0점을 준 경우가 1개, 채점자가 0점을 주었는데 자동채점 프로그램은 1점을 준 경우가 5개이다. 이 문항의 정답은 '평평한' 외에도 '평평한 길'까지 1점을 부여하는데, 채점 불일치는 채점자가 '평평한 길'을 0점 부여했거나 '평탄한'을 1점 부여한 경우로 모두 채점자의 실수에 해당한다. 즉, 단순한 서답형 문항의 경우 자동채점은 정답과의 일치 여부를 정확히 판단하므로 오류 가능성은 거의 없는데, 채점자는 채점 피로도 및 문자 인지의 실수 등으로 일부 오류가 발생한다.

다음 문항 사례로, 중3-5-(1)의 문항에 대해 자동채점 단계별 비율 및 비율 변화를 순서대로 제시한다.

• 문항	• 모범 답안
【서답형 5】 다음 〈자료〉를 읽고 물음에 답하시오. 〈자 료〉 우리를 탈출한 후 열흘 동안 행방이 묘연하던 반달가슴곰 '아름이'가 드디어 잡혔다. 지난 12일 낮 12시 동물원 인근 야산에 숨어 있던 아름이는 열흘 동안 곰의 흔적을 추적해 온 사육사에게 발견되었다. 이어서 동물원에서 지원 팀이 도착하고 한바탕 추격전을 벌인 끝에 ㉠아름이는 사육사에게 생포되었다. 아름이는 동물원에서 건강 검진을 받고 안정을 되찾았다고 한다. (1) ㉠을 '사육사는'을 주어로 하는 문장으로 바꾸어 쓰시오. 사육사는 _____	아름이를 생포하였다.

표 5. 중3-5-(1) 문항의 자동채점 단계별 채점 비율

채점 단계	정답 수	오답 수	채점 수	누적채점 수	미판단 수	채점 비율
모범 답안 일치 채점	1874	129	2003	2003	1007	66.5%
고빈도 답안 일치 채점	68	39	107	2110	900	70.1%

개념 기반 채점	808	1	809	2919	91	97.0%
단서어 기반 오답 처리	0	33	33	**2952**	**58**	**98.1%**
미판단 답안 수작업 채점			58	3010	0	100%

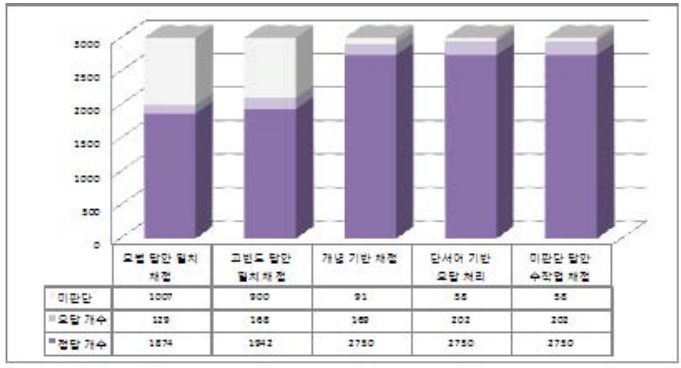

그림 4. 중3-5-(1) 문항의 자동채점 단계별 채점 비율 변화

중3-5-(1) 문항은 2단어 술어형이나 지문상의 주요 단어를 변형하여 답안을 작성하므로 1,000개당 서로 다른 답안 유형 수가 79개로 비교적 응답 자유도가 낮은 편이다. 이런 이유로 정답 패턴으로는 P3 유형에 해당하나 자동채점 가능성은 비교적 높은 문항으로 볼 수 있다.

실제 자동채점 수행 결과, 채점 단계별 채점 비율을 살펴보면 모범 답안 일치 채점 단계(정답 수 1,874, 오답 수 129)에서 약 66%가 채점되고, 이후 고빈도 답안 일치 채점 단계(정답 수 68, 오답 수 39)에서 약 4%, 개념 기반 채점 단계(정답 수 808, 오답 수 1)에서 약 27%, 단서어 기반 오답 처리 단계(정답 수 0, 오답 수 33)에서 약 1%의 채점 비율이 증가하였다. 이는 전체 답안 중 2,952개, 약 98.1%에 해당하는 답안이 채점된 것으로, P3 문항인 것을

고려하면 자동채점 비율이 꽤 높은 편이라고 할 수 있다.

미판단 답안은 58개(약 1.9%)로, 단서어 기반 오답 처리 단계에서 '아름이', '잡았다', '생포' 등의 단어를 포함하는 경우 오답 처리하지 않도록 한 결과, '아름이를 새포했다', '아름이를 생포하이다'와 같이 철자 오류가 있는 경우이거나 '아름이가 다치지않게하기 위해 생포하였다', '도망간아름이를 열흘만에 잡았다'와 같이 정답 외에 다른 단어 · 구를 첨가하여 쓴 경우가 자동채점이 되지 않았다.

채점자와 자동채점 간 결과가 불일치한 답안 수는 10개(약 0.3%)로 채점자가 1점을 주었는데 자동채점 프로그램은 0점을 준 경우가 9개이고, 채점자가 0점을 주었는데 자동채점 프로그램이 1점을 준 경우가 1개이다. 정답은 '아름이를 생포하였다'인데 채점 불일치 사례 중에는 '아름이를 체포했다'와 같이 의미상 혼동하기 쉬운 답안에 대해 채점자별로 각각 다르게 채점한 경우가 있었는데, 자동채점 프로그램은 이를 동일하게 처리하였다.

두 문항의 채점 결과를 자동채점 단계를 중심으로 비교해 보면, P1인 고2-5-㉢ 문항은 고빈도 답안 일치 채점에서 95% 이상의 채점 결과가 완료되고, P3인 중3-5-(1) 문항은 이에 더 나아가 개념 기반 채점에서 97% 이상이 채점 완료가 된다. 즉, P1처럼 간단한 단답형 답안은 모범 답안이나 고빈도 답안과의 일치 여부 정도를 채점자가 판단하여 점수를 부여하면 대다수의 문항이 채점 완료가 되는 데 비해, P3처럼 학생 답안의 응답 자유도가 높아지는 경우는 채점자가 자동채점 시스템이 제공하는 형태소 분석을 통해 학생 답안의 일부 혹은 전체가 정답으로 인정할 수 있는 개념과 일치 혹은 유사한가를 판단하여 점수를 부여하는 단계를 거쳐 단서어 기반 오답 처리 단계를 작동해야 채점 정확도를 유지한 상태에서 채점이 진행되고, 나아가 일부는 자동채점 단계로는 걸러지지 않아 채점자가 수작업 채점을 해야 한다. 〈표 6〉은 다시 두 문항 간 채점 단계별 채점 비율의 양상을 정리하여 보여 준다.

표 6. 사례 문항의 자동채점 단계별 채점 비율 비교

채점 단계	고2-5-ⓒ 문항 1단어 술어형(P1)	중3-5-(1) 문항 2단어 술어형(P3)
모범 답안 일치 채점	75.2%	66.5%
고빈도 답안 일치 채점	**95.7%**	70.1%
개념 기반 채점	95.8%	**97.0%**
단서어 기반 오답 처리	100%	98.1%
미판단 답안 수작업 채점	100%	100%

이 가운데 자동채점의 주요 두 단계인 고빈도 답안 채점 단계와 개념 기반 채점 단계를 사용자 인터페이스로 설명하면 다음과 같다. 〈그림 5〉는 자동채점의 사용자 인터페이스에서 P3인 중3-5-(1) 문항에 대해 고빈도 답안 목록을 생성하여 채점자가 점수를 부여하는 고빈도 답안 일치 채점 단계를 보여 준다. 이 문항의 경우 '고빈도 답안 기준 빈도 설정'을 3으로 한 경우 3,010개의 답안 중 빈도 3 이상인 답안이 11개 유형으로 축약되고 이 11개 답안만 채점자가 점수를 부여하면 누적빈도 '2,886'개가 함께 동시에 채점 완료된다.

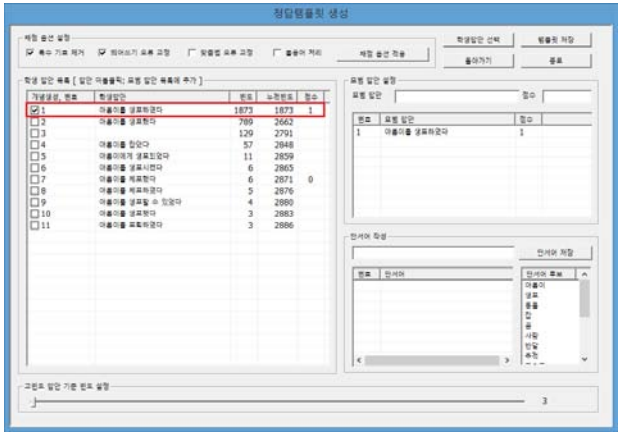

그림 5. 고빈도 답안 채점 단계

모범 답안 일치 채점과 고빈도 답안 일치 채점에서 걸러지지 않은 나머지 답안은 개념 기반 채점 단계로 넘어간다. 개념은 어절 토큰을 단위로 하여 생성 및 편집할 수 있는데, 〈그림 6〉의 '토큰 선택 및 형태소 분석'의 화살표 버튼을 이용해 어절 토큰을 선택하고 형태소 분석 결과를 확인할 수 있다. '유사 토큰 추가'에서 형태소 분석을 통해 어휘 토큰과 문법 토큰으로 분리된 각 토큰에 유사 토큰을 추가할 수 있다. 〈그림 6〉의 예시는 '생포' 대신 '포획'을 추가하여 이렇게 쓴 답안도 정답으로 처리할 것을 기록한 것이다. 여기서 자동채점 시스템이 유의어 목록을 자동으로 추천해 준다.

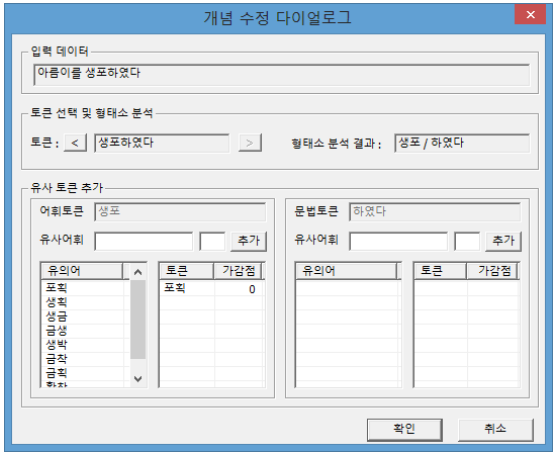

그림 6. 개념 기반 채점 단계

이와 같이 개념 기반 채점 단계에서는 형태소 분석을 통해 다양한 변이 양상을 보이는 유사 답안을 채점할 수 있도록 개념을 추가 또는 편집할 수 있는 도구를 활용하는 것이다.

3. 자동채점 신뢰도 검증

자동채점 프로그램의 채점 신뢰도는 채점자와의 일치도인 Kappa계수, 채점자와의 불일치율로 검증해 볼 수 있다. 따라서 우선 자동채점한 결과를 비교할 수 있는 채점자의 기준점수가 필요하다. 국어 17개 문항을 대상으로 표집한 3,010개의 답안은 먼저 2명의 채점자가 3라운드에 걸쳐 복수 채점하여 최종적으로 일치한 점수를 기준점수로 마련해 둔다. 이렇게 확정된 기준점수와 자동채점 점수 간 Kappa계수를 비교하고, 채점자와 자동채점 간 불일치율[15]을 산출한 결과를 제시하면 〈표 7〉과 같다.

표 7. 국어 문항별 채점 신뢰도 산출 결과

학교급	문항 번호	정답 패턴	사례 수	채점 비율(%)	Kappa 계수	채점 불일치율(%)
초6	2-ⓐ	P1	3,010	100.0	.98	0.07
	2-ⓑ	P1	3,010	100.0	.98	0.07
	2-ⓒ	P1	3,010	100.0	1.00	0.03
	5-ⓐ	P2	3,010	99.44	.99	0.23
	5-ⓑ	P2	3,010	99.93	.99	0.10
중3	1-ⓐ	P2	3,010	100.0	.99	0.56
	1-ⓑ	P2	3,010	100.0	.99	0.50
	5-(1)	P3	3,010	98.07	.97	0.34
	5-(2)	P1	3,010	99.90	.99	0.20
고2	1-ⓐ	P1	3,010	100.0	1.00	0.03
	1-ⓑ	P1	3,010	100.0	1.00	0
	4-(1)-ⓐ	P1	3,010	99.93	1.00	0.23
	4-(1)-ⓑ	P2	3,010	99.93	1.00	0.20
	4-(2)	P4	3,010	98.17	.98	1.02
	5-㉠	P1	3,010	100.0	1.00	0.10
	5-㉡	P1	3,010	100.0	1.00	0.17
	5-㉢	P1	3,010	100.0	1.00	0.20

15 채점 불일치율 = 채점 불일치 수 / (전체 답안 수 - 미판단 답안 수) × 100

P1~P3에 해당하는 단어 · 구 수준 서답형 문항의 Kappa계수는 최소 0.97 이상이며 절반 이상의 문항이 1.00으로 채점 신뢰도가 매우 높다고 할 수 있다. 사실, 자동채점 분야에서 채점 신뢰도는 문항 특성, 채점 기준이나 평가 상황 등에 따라 다르게 해석해야 한다. 본 연구에서 자동채점 프로그램을 적용하고자 하는 학업성취도 평가는 대규모 고부담 시험이므로 엄격한 신뢰도 기준이 요구된다.[16] 또한 적용 대상 문항의 평가 내용이 비교적 명확하고 제한되어 있는 단어 · 구 수준이므로 높은 일치도를 보여야 한다. 이에 일치도를 높이 책정하여 인간채점과 자동채점 간 Kappa계수가 0.80 이상인 경우 채점이 신뢰할 만하다고 해석한다고 가정할 때도, 국어 문항의 자동채점 결과는 전반적으로 채점 신뢰도가 높다고 판정할 수 있다.

또한 P1~P3문항의 채점 불일치율도 0~0.5% 사이로 매우 낮게 나타났다. 자동채점 가능성이 비교적 떨어질 것으로 예상한 P4에 해당하는 고2-4-(2) 문항도 채점 불일치율은 1.02%로 그리 높지 않다. 그런데 이러한 분석 수치상의 우수함에도 불구하고 미판단 답안 및 채점 불일치 사례는 면밀히 검토할 필요가 있다. 우리 교육의 현실에서 대규모 고부담 시험의 경우 하나의 오류라도 인정하기 어려운 것이 현실이기 때문이다. 자동채점이 처리하지 못한 미판단 답안은 정답 외에 다른 단어 · 구를 부가적으로 삽입하거나, 정답에 해당하는 다양한 유의어나 유사 단어가 존재하는 경우로 나타났는데, 특히 답안의 의미 관계를 살펴 부분 점수를 부여해야 하는 경우에 두드러졌다. 이는 현재의 프로그램이 형태소 분석에 그치고 개념 간 유사 정도를 살펴 자동으로 채점하는 단계까지 나아가지 못했기 때문이다. 향후 이 부분에 대한 프로그램의 보완이 요구된다.

---

16 성태제(2002: 394)는 채점자 간 신뢰도를 판단하는 절대적인 기준은 없으나 채점 결과가 점수로 제공될 때는 0.6 이상, 채점 결과가 등급이나 범주로 제시될 때는 일치도 계수 0.85 이상(Kappa계수는 0.75 이상)을 제안하였다. Nehm 외(2012)도 Kappa계수가 0.41~0.60이면 '적정'하고 0.61~0.80이면 '실질'적이며, 0.81~1.00이면 '거의 완전'하다고 평가하였다(Nehm, Ha, & Mayfield, 2012: 187).

종합적으로, 정답 패턴을 기준으로 국어 문항에 대한 자동채점 결과를 정리하면 〈표 8〉과 같다. 검증된 모든 국어 문항의 Kappa계수가 0.97 이상으로 자동채점 결과는 신뢰할 만하다고 하겠다. 다만, 요구하는 답안이 길고 복잡해질수록, 즉 P1에서 P4로 갈수록 인간채점과 자동채점 결과의 일치도가 떨어지고 불일치율이 증가한다는 점은 확인할 수 있다.

표 8. 국어 문항의 정답 패턴별 채점 결과(평균)

교과	정답 패턴	채점 비율(%)	Kappa계수	채점 불일치율(%)
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

IV. 결론

한국어 서답형 자동채점 프로그램은 P1~P3 비중이 높은 학업성취도 평가의 단어 · 구 수준 문항에서는 자동채점이 가능한 것으로 나타났다. 즉, 자동채점 프로그램의 채점 비율과 채점자와의 일치도는 적정한 수준이다. 다만, 일부 미판단 답안이나 불일치 답안을 볼 때 유사어 처리 문제와 다른 문자나 용어가 포함되어 있는 경우 등을 고려하여 프로그램을 보완할 필요가 있다.

한국어 처리 기술 및 지식베이스의 여건이 충분하지 않은 상황에서, 자동채점이 용이한 단답형 문항부터 자동채점 시스템을 구축하는 것은 출발점으로서의 의미가 크다. 현재 대규모 평가의 서답형 문항에서 다수를 차지하고 있는 단답형 문항을 우선 처리할 수 있을 뿐 아니라, 이를 바탕으로 좀 더

장기적으로 내용 기반의 서답형 문항에 대한 자동채점 연구도 지속적으로 발전시킬 수 있는 기반을 마련할 수 있기 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.

* 본 논문은 2014. 1. 31. 투고되었으며, 2014. 2. 7. 심사가 시작되어 2014. 2. 28. 심사가 종료되었음.

참고문헌

강원석(2011),「질의문 유형 분석을 통한 서답형 자동채점 시스템」,『한국콘텐츠학회논문지』 11(2), pp. 13-21.

교육과학기술부(2009. 12),『2009 개정 교육과정 총론』(교육과학기술부 고시 제 2009-41호).

_____(2010. 5),「창의성과 인성 함양을 위한 교육내용 · 방법 · 평가체제 혁신 방안 VIP 보고」, 대통령 주재 제3차 교육개혁 대책회의(2010년 5월 19일).

_____(2011. 3),「2011 창의 · 인성 교육 기본 계획. 보도자료」(2011년 3월 11일).

_____(2011. 6),「스마트교육 추진 전략(안)」, 보도자료(2011년 6월 29일).

_____(2011. 12),「창의 · 인성교육 강화를 위한 중등학교 학사관리 선진화 방안 발표 — 고교 석차 9등급제 평가를 성취평가제로 전환 —」, 보도자료(2011년 12월 14일).

권오영(2004),『웹 기반 주관식 평가문항 채점 알고리즘 설계 및 구현』, 한서대학교 교육대학원 석사학위논문.

김경성 · 김종훈 · 곽현석(2011),『2012년 국가수준 학업성취도 평가 채점 표준화 방안 연구』(교육과학기술부 수탁과제 2011-1), 서울: 한국교육과정평가원.

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

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

박희정 · 강원석(2003),「유의어 사전을 이용한 주관식 문제 채점 시스템 설계 및 구현」,『한국컴퓨터교육학회논문지』 6(3), 207-216.

성태제(2002),『현대교육평가』, 서울: 학지사.

성태제 · 양길석 · 강태훈 · 정은영(2010),『학업성취도 평가 서답형 문항 컴퓨터 채점화 방안 연구』, 한국교육과정평가원 연구보고 CRE 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.

송미영 · 박혜영 · 임해미 · 최혁준(2013),「21세기 역량 평가를 위한 OECD PISA의 변화 방향과 대응 방안」(2013 KICE 이슈페이퍼 / 연구자료 ORM 2013-57-13), 서울:

한국교육과정평가원.
정동경(2001),『벡터 유사도와 시소러스를 이용한 주관식 답안의 채점 방법』, 동국대학교 교육대학원 석사학위논문.
조우진(2006),『의미 커널과 한글 워드넷에 기반한 지능형 채점 시스템』, 한림대학교 대학원 석사학위논문.
한국교육평가학회(편)(2004),『교육평가용어사전』, 서울: 학지사.
Dube, T. & Ma, M.(2010), "Marking short free text responses in e-assessment," Retrieved from http://www.heacademy.ac.uk/assets/ocuments/subjects/ics/eteaching_marking_short_free_text_responses.pdf. 16-18.
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.

초록

국가수준 학업성취도 평가 국어 서답형 문항의 자동채점 결과 분석

노은희

이 연구는 2013년 개발된 한국어 서답형 문항 자동채점 프로그램 KASS(Korean Automatic Scoring System)을 활용하여 2012년 학업성취도 평가의 초 · 중 · 고 국어 총 17문항 각 3,010개 답안을 대상으로 자동채점 결과를 분석하여 향후 국어 서답형 문항의 자동채점 가능성을 탐색하였다.

자동채점 결과, P1~P3에 해당하는 단어 · 구 수준 서답형 문항의 Kappa계수는 최소 0.97 이상이며 절반 이상의 문항이 1.00으로 채점 신뢰도가 매우 높게 나타났다. 또한 인간채점과 채점 불일치율도 0~0.5% 사이로 매우 낮게 나타났다. 다만, 요구하는 답안이 길고 복잡해질수록 인간채점과 자동채점 결과의 일치도가 떨어지고 불일치율이 증가한다는 점은 확인할 수 있다.

한국어 처리 기술 및 지식베이스의 여건이 충분하지 않은 상황에서, 자동채점이 용이한 단답형 문항부터 자동채점 시스템을 구축하는 본 연구는 출발점으로서의 의미가 크다. 현재 대규모 평가의 서답형 문항에서 다수를 차지하고 있는 단답형 문항을 우선 처리할 수 있을 뿐 아니라, 이를 바탕으로 좀 더 장기적으로 내용 기반의 서답형 문항에 대한 자동채점 연구도 지속적으로 발전시킬 수 있는 기반을 마련할 수 있기 때문이다.

핵심어 자동채점, 한국어 자동채점 프로그램, 서답형 문항, 학업성취도 평가

ABSTRACT

Application of an Automatic Scoring Program for Short Answer of Korean Items in NAEA

Noh, Eun-hee

The purpose of this study is to improve the automatic scoring program of Korean supply-type items and to increase application of the program for effectiveness of scoring and reliable scoring. For the trial application of the automatic scoring program, I scored 17 supply-type Korean items of the 2012 NAEA(National Assessments of Educational Achievement). The numbers of answer were 3,010 of each Korean item. The results of the 2012 NAEA items demonstrated that the scoring rate was quite high, 97~100%, and Kappa coefficients were high (at over .97). The rate of scoring errors was 0~0.5%. The error rate of most items was very small (under 1%). The sources of scoring errors were either spelling errors or the non-recognition of analogous terms, and symbols.

The ability to use automatic scoring program in operational scoring environments, such as the NAEA, reduces the time and cost associated with having multiple human scorers score answers of supply-type items. Therefore, an automatic scoring would appear to be a favorable solution with respect of both the introduction of more supply-type items on high-stakes standardized tests and on the lower stakes classroom-instruction environment.

KEYWORDS automatic scoring, Korean automatic scoring program, supply-type items, National Assessments of Educational Achievement