영어 받아쓰기에 대한 전산채점과 수기채점의 결과 비교 분석*

이보림** · **최종헌**
(원광대학교 · 성남영어마을)

Lee, Borim & Choi, Jonghurn. (2014). Comparative Analyses of Different Scoring Methods on English Dictation Tests. *The Linguistic Association of Korea Journal, 22*(4), 339-358. This study discusses the scoring methods to assess Korean students taking an English dictation test. We compared three different dictation assessment results from 21 native English teachers, 8 Korean English teachers, and a computer, all of which assessed 30 Korean sixth graders. We also tried to figure out how students' errors defined by linguistic categories are related to the differences in scoring and attempted to discover what factors have caused them. The results showed distinct differences between the scores given by native English teachers and those given by Korean teachers, with the computerized scores almost always being the lowest. The results can contribute to a better algorithm for scoring, so that automatic instant scores that most teachers would agree on can be given by computer. It will save time and provide students with better suggestions for what they would need to focus on in their future studies.

주제어(Key Words): 받아쓰기(dictation), 전산채점(computer-based scoring), 수기채점(human scoring), 오류분석(error analysis)

1. 서론

받아쓰기는 언어 학습의 가장 기초적인 도구이다. 또한 선택형 답과 달리 서술형이라는 점에서, 그리고 일반 주관식 서술형과 달리 정답이 단 하나인 객관식 서술형이란 점에서 학

* 이 논문은 2012학년도 원광대학교의 교비지원에 의해서 수행 됨. 유익한 논평을 통해 논문의 내용을 충실하게 만드는데 큰 도움을 주신 익명의 심사자께 감사드린다.

** 제1저자 겸 교신저자

생이 무엇을 잘못 들었는지 또는 모르는지를 평가할 수 있는 유용한 도구이다. 받아쓰기의 채점에 관해서는 총 몇 단어 중 몇 개 틀렸다는 식의 단순 채점이 일반적인 채점 형태였다. 하지만 똑같은 개수의 오류일지라도 해당 오류가 단순한 철자의 오류인지, 음운적, 형태적, 통사적 오류인지, 또는 의미적으로 문장의 뜻을 아예 반대로 만드는 오류인지를 구별해서 평가한다면 보다 바람직한 효과를 기대할 수 있을 것이다. 즉, 해당 학습자의 학습 상황에 대한 보다 적확한 평가가 이루어질 수 있으며, 평가 후 학습자 본인이 신경을 써서 학습해야 할 부분을 인지함으로써 보다 효율적인 학습효과를 성취할 수 있을 것이다.

Oller(1972)에 의하면 받아쓰기(dictation)는 일반적으로 교사에 의해 제공되는 발화문을 귀로 인지하고 그 내용을 머리에서 보존하여 그렸다가 문자로 재구성해 내는 종합적인 과정을 포함한다고 한다. Postovsky(1974)는 듣고 쓰는 과정인 받아쓰기를 통하여 언어 습득의 네 가지 기능인 듣기, 쓰기, 읽기, 말하기 능력이 쉽게 개발될 수 있다고 주장했는데 이는 받아쓰기의 긍정적인 전이효과를 의미한다. Oller(1979)도 역시 받아쓰기는 여러 언어 능력, 특히 듣기 능력 향상을 위한 유용한 방법이라고 주장했다. 특히 1970년대 이후 언어 능력 평가 시 종합적인 의사소통에 더 중점을 두게 되면서 외국어 능력 측정의 도구로서 받아쓰기의 유용성이 크게 부각되었다(Rivers & Temperly, 1981).

듣기 평가에 관한 많은 연구가 있으나(Buck 2001 참조) 본 연구에서는 듣고 쓰는 과정을 포함하는 받아쓰기 평가만을 다루기로 한다. 받아쓰기 평가는 주로 부분 받아쓰기(spot dictation)와 전체 받아쓰기(full dictation)의 두 유형으로 나뉘는데(Paulston & Bruder, 1976), 전체 받아쓰기는 학습자가 청취한 내용을 모두 받아쓰는 형식으로서 학습자들의 언어 습득 과정을 종합적으로 평가할 수 있어서 유용한 평가 방식이다. 반면에 전체 받아쓰기 답안에 대한 평가 또는 채점은 받아쓰기 자체가 한 개의 정답을 가진 서술형 문제라는 점에서 채점 작업이 복잡하고 객관적 평가가 용이하지 않다. 채점결과의 객관성에 관해서는 평가자 간의 신뢰도(inter-rater reliability)와 평가자 내의 신뢰도(intra-rater reliability) 문제가 중요하게 지적된다(McNamara & Candlin, 1996).

따라서 받아쓰기 평가나 채점에 관한 연구는 상대적으로 매우 빈약한 상황이다. 김영철(1988)은 중학생들의 받아쓰기 평가절차에 대한 연구에서 채점상의 어려움을 해결한 보다 체계적인 받아쓰기 평가절차를 제시하고자 하였다. 정소라(2008)는 받아쓰기에 대한 네 가지채점방식을 비교하고 의미의 정확도 여부에 따라 정확히 맞으면 1점, 틀리면 0점을 주는 방식이 가장 효과적임을 밝혔다. 이선주(2006)는 받아쓰기 평가는 종합테스트의 훌륭한 예시이므로 멀티미디어를 활용했을 때 특히 그 장점이 부각된다는 주장을 하였다.

받아쓰기 외의 분야에서는 주로 작문 채점에 대한 연구들이 수행되었다. 작문수행평가에 대한 전산채점의 활용 가능성에 대한 연구에서 언어능력 평가 전문가와 전산채점 결과를 비교하여 전산채점의 신뢰성을 조사하였는데, 전산채점이 요구되는 신뢰성을 충족시키기 위해서는 좀 더 발전된 채점 규칙이 개발되어야 함이 지적되었다(최인철과 엄연희, 2001; 최인철,

임해창, 박정, 2003). 또한 최윤정(2006)은 외국에서 개발된 여덟 종류의 에세이 채점 프로그램들에 대한 공통점과 차이점을 비교하고 문제점과 한계점을 분석하였다. 그리고 김보라와 이규민(2012)은 초등학교 쓰기 수행평가에 대한 채점 방식에서 점수에 영향을 미치는 요인들의 상대적 영향력을 산출하고, 잘 정의된 채점기준과 절차 및 채점자들의 훈련이 쓰기 평가 채점의 신뢰도를 높이는데 기여할 것이라고 밝혔다.

오류분석에 관한 연구인 김원명(1984)은 고등학생들의 영어 청해 능력을 파악하기 위해 영어 청취상의 오류를 수집하고 분석하였다. 이현구와 윤병남(2007)은 중학생 영어학습자의 영어 받아쓰기에서 음운인식, 연음 등의 구체적인 음운적 오류분석과 음운 오류들 간의 난이도를 분석하였으며, 받아쓰기에서 나타난 형태적, 의미적, 통사적 오류분석에 대한 연구가 필요하다고 지적하였다.

그러나 받아쓰기에 나타난 언어학적 오류들과 채점방식 간의 연관성에 대한 연구는 전혀 찾아볼 수 없었다.[1] 본 연구는 영어 받아쓰기 답안을 평가하기 위한 채점방식의 연구이다. 평가의 목적은 평가를 받는 대상자(학습자)의 현재 학습수준과 문제점(약점)을 정확히 파악하는 것인데, 학습자의 문제점은 시험에서 오답으로 나타나므로 받아쓰기 채점에 있어서도 오답들, 즉 오류들을 파악하여 어떻게 점수에 반영을 하느냐가 핵심이 될 것이다. 전산화된 채점이 이상적이기는 하지만 수기채점에서 반영하는 모든 요인들을 점수화하여 전산화 하는 작업은 쉬운 일이 아니며 인공 지능이 제대로 채점하는 날이 올 때까지는 100% 만족하는 채점 시스템을 가지기는 불가능한 일일 것이다.

이상적인 받아쓰기 채점 알고리즘을 개발하기 위해서는 이론에 근거한 것이 아닌 수기채점의 결과물에 가장 가깝게 채점할 수 있는 현실적 공식이 필요할 것으로 보인다. 따라서 본 연구의 목적은 영어 교사들의 수기 채점의 결과들을 여러 가지 언어학적 오류 요인들에 비추어 비교 분석함으로써 영어 교사들이 임의로 사용했을 수기 채점의 평가 지표들을 추출하여 분석하고 각각의 중요도를 비교 분석해 보는 것이다.

2. 연구 방법

받아쓰기 채점의 바람직한 알고리즘을 개발하기 위한 하나의 기반 연구로서 본 연구에서는 경기도 S시에 소재한 S영어마을에서 시행하는 영어 받아쓰기 시험의 채점에 사용되는 전산채점 방식과 영어 교사들이 수기로 한 채점 결과들을 비교 분석하였다. 수기채점은 영어 원어민 교사들과 한국인 영어 교사들의 두 집단에 의해 시행되었고, 전산채점과 두 가지 수

1) 본 연구에서는 받아쓰기의 정답과 일치하지 않는 모든 양상을 '오류'로 규정하여 분석하였으며, 이러한 오류들을 여러 각도로 분석하는 것을 '오류분석'이라고 규정하여 사용하였다.

기채점의 결과들을 본 연구에서 사용한 12가지의 언어학적 오류들을 기준으로 채점 방식 간의 차이점과 연관성을 비교 분석하는 방법을 사용하였다.

2.1. 연구 대상

본 연구는 S영어마을에서 2011년도에 수도권 초등학교 학생들에게 실시한 받아쓰기 평가시험을 치른 300여명의 6학년 학생들 중에서 E사에서 개발한 전산채점 방식에 의한 평가결과 가장 높은 점수를 받은 30명의 답안을 대상으로 하였다. 상위 30명만을 대상으로 한 이유는 본 연구가 피험자들의 오답과 그 오답에 대한 평가에 관한 연구이기 때문에 너무 터무니없는 오답들을 배제하여 연구 결과의 신뢰도를 높이기 위한 것이었다.

수기채점자들은 경기도 S시에 소재한 S영어마을 소속의 원어민 교사 21명과 한국인 교사 8명이 자원하여 연구에 참여했다. 원어민 교사들의 국적은 미국 12명, 캐나다 4명, 영국 1명, 호주 1명, 그리고 남아프리카공화국 국적의 백인이 3명이다. 그 중 18명은 학사, 3명은 석사 학위 소지자이며, TESOL/TEFL 자격증 소지자는 11명이었다. 한국인 교사들 역시 학사 학위 이상의 학력을 가졌고, 그 중 석사학위 소지자가 2명이었으며 TESOL/TEFL 자격증 소지자는 2명이었다. 원어민들 중 4명만이 언어학, 영문학 분야 전공자였으며, 한국인 중에는 5명이 그 분야의 학위를 소지하고 있었다. 그러나 모든 채점자들은 학교와 학원 등에서 영어를 상당 기간 동안 가르친 경력을 가진 자들이다. 연구가 진행될 당시를 기점으로 원어민 교사들의 영어 교사 경력 기간은 최소 1년에서 7년까지로서 평균 2년 8개월이었으며, 한국인 교사들은 최소 1년에서 8년까지로서 평균 3년 5개월이었다.

해당 평가 시험은 E사가 개발한 1급부터 9급까지의 영어 받아쓰기 등급 중에서 초등학생에게는 가장 높은 수준인 5급 시험으로서 기초 확립 수준이라고 볼 수 있다(에듀조선영어연구소, 2010). 5급 시험은 교과부 고시 기초 735단어 수준을 약간 상회하는 약 800 단어 내에서 출제된다. 5급 통과를 위하여 반드시 알아야 하는 75가지 문형을 정해놓고 그 안에서 1-4형식 문형을 이용한 단문과 의문문으로 구성되어있다. 평가는 총 네 개의 파트로 구성되어 있고 총 문항 수는 100개이나 그 중 파트 3의 전체 문장 받아쓰기 문항인 총 25문항만을 분석 대상으로 정하였다. 대상 문항들은 51번부터 75번까지 최소 4단어에서 최대 7단어로 구성되었고, 각 문항의 정답은 표 1에 제시하였다.

평가 원문 녹음은 미국 원어민 강사가 정상 발화 속도로 녹음한 것을 사용하였고, 각 문항은 총 3회씩 들려주었는데 처음 두 번의 녹음 재생 후에는 6초 동안 쓸 시간이 주어졌고 마지막 세 번째 재생 후에는 12초의 시간이 주어졌다.

표 1. 문항 정답

문항번호	문장
51	Please do your work.
52	Wash your hands before dinner.
52	He played the piano well.
54	I have finished my work.
55	I have been studying English.
56	We should finish it by tomorrow.
57	I would like to have some food.
58	I'll finish it by tomorrow morning.
59	You must have rained.
60	It must have rained.
61	You must listen to your mother.
62	I could go there later.
63	You'd better stop eating too much.
64	She has to go there.
65	I can make it there at six.
66	Can you get me some soda?
67	Would you like to have some pizza?
68	Will you please give me a call?
69	Could you write your address there?
70	Are you going to be there?
71	I won't go there tomorrow.
72	I can't find my English book.
73	You don't have to do that.
74	I couldn't do it there.
75	I wouldn't talk to him.

2.2. 전산채점 방식

E사에서는 학습자들이 스스로 학습할 수 있는 온라인 단계별 영어 받아쓰기 시스템을 개발하면서 시간이 걸리지 않고 거의 무료로 이용할 수 있는 자동채점 방식을 개발하였다. 그 당시 온라인 시스템에서 사용되는 자동채점 방식은 순서에 관계없이 몇 개의 단어를 맞추었는지를 보거나, 각 단어의 정확한 순서에 특정 단어가 있느냐를 따지는 정도였다. 이러한 문제점들을 보완하기 위하여 E사에서는 컴퓨터에 수학적 공식을 사용하는 프로그램을 이용하는 방식으로 오류를 산출하는 자동 채점방식으로 점수를 냈는데 그 기준은 다음과 같다.

채점 공식은 총 다섯 가지로서 각 공식 당 1점이 배점된 후 다섯 가지 항목의 합에다 곱하기 1.2를 하여 각 문항은 6점 만점으로 계산된다. 첫째 공식은 단어 수의 일치를 평가한다. 예를 들어 단어 5개로 구성된 문장일 경우, 답안으로 입력한 단어 수가 5개로 일치하면 1점,

한 단어의 차이(입력 단어수가 4개나 6개일 경우)가 나면 1/n만큼 감점하므로 1/5을 감점하여 0.8점이 되고 두 단어의 차이(입력 단어수가 3개나 7개)를 보이면 0.6점, 세 단어의 차이(입력 단어수가 2개나 8개)가 생기면 0.4점, 네 단어 차이(입력 단어수가 1개나 9개)가 나면 0.2점을 받게 되고 입력 단어수가 없거나 10개 이상인 경우에는 절대값 기준으로 0점 이하로 내려갈 수 없으므로 0접이 배점된다. 두 번째 공식은 문장에서 정답과 첫 단어와 끝 단어가 일치하느냐로 보아서 각각의 경우에 0.5점이 배점되어 첫 단어와 마지막 단어 둘 다 맞으면 1점, 하나만 맞으면 0.5점, 둘 다 틀리면 0점이 된다.

세 번째 공식은 단어의 철자까지 정확한 일치 여부를 평가하는데 순서는 관계없이 정답에 포함된 단어가 정확히 입력 답안에도 있는가를 본다. 4개의 단어로 구성된 문항의 예를 들어 설명한다면 정답의 단어 4개가 모두 있으면 1점이지만 정확히 매치되는 단어가 3개라면 1/n(1/4)이 감점되어 0.75점, 2개가 매치되면 0.5점, 한 개뿐이라면 0.25점, 그리고 한 개도 없다면 0점이 된다. 네 번째 공식은 정답의 단어를 두 개씩 묶어서 묶여진 단어 쌍들이 순서대로 되어있는지를 보고 그 쌍의 수가 맞는지를 평가한다. [Please do your work.]라는 51번 문항을 예로 들어 설명하면 단어를 순서대로 <Please do>, <do your>, <your work>의 세 쌍이 생성되고 이 경우 n의 값은 3이 된다. 이 공식의 경우 n 값은 총 단어 수의 -1이 된다. 입력된 답안이 [Please do your work please.]라면 <Please do>, <do your>, <your work>, <work please>의 네 쌍이 생성되고 세 쌍이 정답과 일치하므로 만점일 것 같지만 <work please>라는 정답에는 없는 쌍이 하나 있으므로 1/3이 감점되어 2/3점만 받는다. 또한 답안에 [Please do]만 입력되었다면 <Please do> 한 쌍이 일치하여 1/3점만 받는다. 또 다른 예로 [Please do whatever]라는 오답은 <Please do> 한 쌍은 맞지만 <do whatever>라는 정답에 없는 쌍이 또 하나 있으므로 0점이 된다. 물론 틀린 쌍이 맞는 쌍의 수보다 많아도 최소 점수인 0점 이하로 내려가지는 않는다. 마지막으로 다섯 번째 공식은 정답의 단어를 세 개씩 묶어서 묶은 단어 그룹들이 순서대로 되어 있는지를 보고 맞는 수에 따라 배점한다. 위에서 본 것과 동일한 문항을 예로 들어 설명하면 [Please do your work.]는 <Please do your>와 <do your work>의 두 단어 그룹이 생성되므로 n 값은 2가 된다. [Please do your work please.]라는 오답은 <Please do your>, <do your work>, <your work please>의 세 그룹이 생성되어 앞의 두 그룹은 일치하므로 0.5x2가 되어 1점 만점이 될 것 같지만 마지막 그룹이 틀렸으므로 다시 0.5점이 감점되어 결과적으로 0.5점만 받게 된다. [Please do your whatever]라는 오답은 <Please do your> 때문에 0.5점을 받지만 <do your whatever> 때문에 다시 0.5점이 감점되어 결국 0점이 된다.

2.3. 언어학적 오류 유형

본 연구에서는 또한 언어학적 범주(linguistic categories)를 기본으로 하는 오류를 산출하여 분석하여 보았다.[2] 이현구와 윤병남(2007)은 Brown(1980)의 수학적 범주와 언어학적 범주에 기본을 두고 음소인식, 연음, 기능어, 축약/약화/탈락/첨가 및 철자법 등 다섯 가지 오류를 분석한 바 있다. 본 연구에서는 표 2에 제시한 총 12가지의 언어학적 오류를 다루었다.

철자법 오류는 틀린 철자를 가진 단어의 개수를 통계로 냈는데 예를 들어 'address'를 'adres'로 쓰는 식이다. 형태적(morphological) 오류에는 단어 탈락, 단어 첨가와 단어 부조화(mismatch) 오류를 분석하였고, 대상 단어가 내용어인지 기능어인지에 따라 오류의 강도가 달라질 수 있다고 보았기 때문에 그것도 따져보았다. 단어 탈락이나 첨가의 경우는 설명이 필요 없으므로, 단어 부적합 오류의 예만 들자면, 'too' 대신 'to'를 'won't' 대신 'will'을 쓴 경우를 의미한다.

통사적(syntactic) 오류로는 접미사 탈락, 첨가, 부조화의 세 가지 오류를 보았다. 예로는 'finished' 대신 'finish'를 쓰거나 그 반대의 경우가 접미사 탈락과 첨가의 경우이며, 'played' 자리에 'plays' 또는 'playing' 등을 쓴 경우가 접미사 부조화 오류의 예이다. 의미적(semantic) 오류에는 부정을 긍정으로 쓴 오류가 있는데 주로 조동사와 'not'의 축약형에서 오류가 생겼다. 즉, 'won't'를 'will'로 'can't'를 'can'으로 잘못 알아듣고 쓰는 경우이다. 마지막으로 합체 오류와 분리 오류가 있는데 이 두 가지는 음운적(phonological) 오류라고 볼 수 있다. 우선 합체 오류는 주로 단어 사이에 연음이 되는 것을 파악하지 못하고 같은 단어의 일부로 잘못 알아들은 경우가 대부분이었다. 예를 들면 'finish it'를 'finished'로, 'have been'을 'having'으로, 그리고 'have rained'를 'brained'나 'bringed'로 쓰는 경우이다. 반면에 분리 오류는 축약된 한 단어를 두 단어로 들어서 쓰거나 접미사를 다른 단어로 잘못 듣고 쓰는 경우들이다. 예를 들자면 'I'll'을 'I will'로 떼어 쓰거나 'studying'를 'study in'의 두 단어로 틀리게 쓰는 경우이다.

---

2) Brown(1980)은 수학적 범주의 오류와 언어학적 범주의 오류를 구분하였고 형성 단계별로 체계화 이전단계와 체계화 단계, 그리고 체계화 이후 단계로 오류의 범주를 기술하였다. Duskova(1969)는 형태, 서법동사, 시제, 관사, 어순, 동사, 전치사, 구조, 어휘 등의 문법 분류 방식을 이용하여 아홉 가지로 오류를 분류하였다. 반면에 James(1977)은 관사, 시제, 어순, 어휘, 부정사, 변형, 주어 동사의 일치 등 일곱 가지로 분류하였다.

표 2. 언어학적 오류의 종류

오류	설명
Spell	철자법 오류
Sf-d	접미사(suffix) 탈락 오류
Sf-a	접미사(suffix) 첨가 오류
Sf-mm	접미사(suffix) 부조화(mismatch) 오류
Wd-d	단어 탈락 오류
Wd-a	단어 첨가 오류
Wd-mm	단어 부조화 오류
Func	기능어 오류
Cont	내용어 오류
Neg-Pos	부정-긍정 교체 오류
Coal	합체(coalescence) 오류
Div	분리(division) 오류

3. 연구 결과 분석

세 가지 채점방식들의 결과들을 비교하기에 앞서 채점방식들 내의 기본적인 일치도를 보기위해 켄달의 타우(Kendall's τ)를 계산하여 그 결과를 3.1에 제시하였다. 채점 결과는 채점방식별로 주목할 만한 점수 차이가 있는 경우들을 모아서 3.2에서 다루었으며, 언어학적 오류의 종류를 12가지로 분석하여 각 답안에 해당하는 오류를 중복 체크하여 문항별로 분석한 결과를 3.3에서 다루었다. 마지막으로 언어학적 오류들을 기준으로 하여 세 가지 채점방식별 평가결과들을 통계적 분석 방법을 이용하여 비교하였다. 분석에 사용한 통계 방법은 Friedman 검정과 Bonferroni 방법을 이용한 사후분석이며, 3.4에 자세한 결과들과 그 결과들이 의미하는 바를 제시하였다.

3.1. 채점방식들 간의 신뢰도 및 평균 순위 비교

수기채점 방식의 채점자들은 원어민 21명과 한국인 8명이었는데, 각 채점자는 각 문항을 6점 만점으로 채점하도록 요청받았다. 다만 대문자 소문자 오용에 대한 감점은 하지 말 것을 주지하였다. 본 연구의 목표가 수기 채점자들이 각자의 기준을 가지고 모두 성실하게 채점을 하였다는 가정 하에, 채점자들이 피험자들의 오류별로 어느 정도의 비중을 두고 감점을 하는가를 역으로 분석하여 채점자 그룹 간의 차이를 비교 분석하는 것이었기 때문에 채점자들에

게 채점의 구체적인 기준들을 제시하지 않았다.

명확한 기준이 없는 상태에서 채점을 했으나 모든 채점자들이 동일한 개념을 측정하였기 때문에 그 결과가 비슷하게 나타나야 한다는 것을 전제로 원어민 그룹 내의 21명과 한국인 그룹 내의 8명의 채점자들이 각각 얼마나 일치되게 채점을 하였는지 알아보고자 하였다. 즉 채점자의 신뢰도를 알아보기 위하여 각 문항에 대하여 6점 만점으로 채점한 점수를 서열형 척도로 간주하여 채점자 간의 일치도를 살펴보기 위해서 비모수 상관계수 중 하나인 켄달의 타우를 계산하였다.[3] 켄달의 타우값은 최솟값 0.79부터 최댓값 0.97까지로 나타났고, 평균값은 원어민 간에는 0.9207, 한국인 간에는 0.9017로 나타났다. 타우값이 0.8 이상이면 일치하는 것으로 간주되므로 이 결과는 상당히 신뢰할만한 일치도를 나타내는 값이라고 할 수 있다.

표 3. 수기채점자들의 신뢰도 분석

채점자	짝의 개수	Kendall's τ			
		평균	표준편차	최솟값	최댓값
원어민 간	210	.9207	.01968	.87	.97
한국인 간	28	.9017	.04312	.79	.96

또한 본 연구에서는 각 채점자들의 수가 적기 때문에 데이터가 매우 적은데 정규성이 의심되거나 데이터에 이상이 있는 경우 사용하는 분석 방법인 Friedman 검증을 실시하였다. 우선 원어민 채점자와 한국인 채점자들의 점수와 전산 점수 간의 결과를 비교하기 위해 각 채점방식의 평균과 표준편차, 채점자와 전산 점수 간의 차이에 대해 Friedman 검증을 한 결과는 표 4와 같다.

표 4. 채점방식 간의 점수 비교

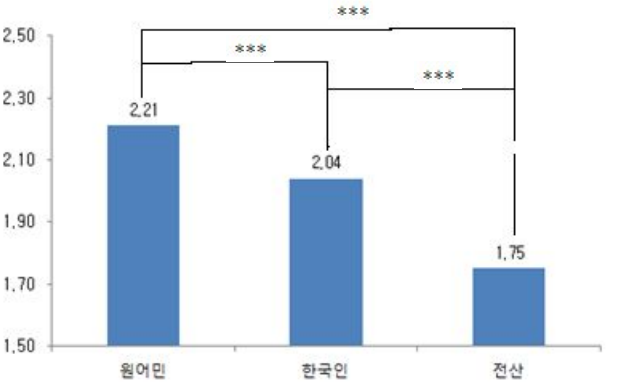
채점방식	평균	표준편차	평균순위	df	χ^2
원어민	5.48	0.80	2.21		
한국인	5.43	0.85	2.04	2	222.520***
전산	5.21	1.14	1.75		

***p<.001

표4에서 보듯이 수기채점자들과 전산채점 점수 간에 통계적으로 유의한 차이를 보였다(χ^2=222.520, p<001). 점수를 나열한 뒤 점수에 순위를 매겨서 순위로 평균을 내는 평균 순위

3) 켄달의 타우는 동일 점수부여가 가능한 서열척도의 경우에 적용되는 비모수 상관계수이다.

를 보면 외국인 채점자 그룹이 2.21로 가장 높고, 그 다음으로 한국인 채점자 그룹이 2.04, 그리고 전산채점방식은 1.75로 가장 낮은 점수가 나왔음을 알 수 있다. 사후분석으로는 Bonferroni 방법을 사용하였고, 보정된 유의수준은 0.017로 그림 1에 표시된 바와 같이 전산과 원어민, 전산과 한국인, 그리고 원어민과 한국인 간의 모든 경우에 점수 차이가 유의한 것으로 나타났다.



보정된 유의수준 =0.017 by Bonferroni's method
그림 1. 채점방식들 간의 평균 순위

3.2. 채점방식들 간의 점수 차이 분석

본 연구의 분석 대상이 되는 토큰수는 25문항에 대한 30명의 수험자가 있으므로 총 750개가 된다. 채점방식 간에 큰 점수 격차를 보이는 토큰들을 각각 15개씩 모아서 점수 차이를 보고 그 이유를 분석해 보았다.

우선 수기채점(원어민과 한국인을 통합) 점수가 전산채점 점수보다 높은 토큰 15개를 추려보니 6점 만점에서 두 채점 방식 사이의 점수 차이가 2.77점에서 1.86까지 벌어졌다. 가장 격차가 큰 토큰은 62번의 [I could go later.]에 대한 오답 [I'd go later.]로서 수기채점자들은 내용상 의미가 거의 보존되었기에 평균 4.57의 점수를 주었으나 전산채점의 경우에는 공식에 근거하여 감점이 많았으므로 불과 1.8점이라는 점수가 부여되었다. 그 다음으로는 52번 문항 [Wash your hands before dinner.]에 대한 오답 [wash your hand before dinner]로서 수기점수는 5.43점이고 전산점수는 2.82점이었다. 중간에 나오는 단어의 복수 접미사 하나가 빠졌을 뿐인데 전산에서는 연결된 단어들(2쌍, 3쌍)에서 계속 감점이 되는 결과가 생기므로 과중치가 높아져서 전산 점수가 크게 깎이는 결과를 빚게 되었다. 전체적으로

볼 때 전산에서는 중간에 들어간 단어에 대한 작은 실수나 상대적으로 덜 중요한 단어가 하나 빠졌을 때에 과중한 감점이 됨을 알 수 있었다.

반대로 전산 점수가 수기 점수보다 높은 토큰 15개를 추려보았더니 점수 차이가 1.03점에서 0.36점으로 나타났다. 이 부류의 가장 큰 공통점은 중요한 내용어에 오류가 있는 경우로서 'won't'가 'want'나 'went'로 쓰인 경우이거나 'rained'와 같은 본동사가 누락된 경우들이었다. 즉 주요 단어의 오류가 있거나 탈락이 될 경우 수기채점자들은 감점을 많이 했으나 전산채점에서는 단어 한 개의 실수로 상대적으로 덜 중요한 기능어의 오류와 동일하게 처리되는 문제가 부각되었다.

다음에는 수기채점자 그룹을 원어민과 한국인으로 나누어 점수 차이가 큰 토큰들을 15개씩 추려서 비교해 보았다. 원어민 점수가 한국인 점수보다 높은 경우의 점수 차이는 최대 1.2점에서 최소 0.8점을 보였다. 가장 큰 점수 차를 보인 토큰은 71번 문항의 [I won't go there tomorrow.]를 [I would go there tomorrow.]로 쓴 경우였는데 원어민 점수 평균은 5.1점이었는데 반하여 한국인 평균은 3.9점이었다. 69번 문항의 'write'를 'right'로 쓴 경우에도 원어민 그룹은 한국인 그룹에 비해 평균 0.9점 높은 점수를 주었다. 종합적으로 볼 때, 원어민은 비슷한 발음의 다른 단어를 쓴 경우 상대적으로 높은 점수를 준 반면에 한국인은 이러한 오답을 심각한 오류로 여기는 경향이 있었다.

반대로 한국인 채점자 그룹이 원어민 그룹에 비해 높은 점수를 준 토큰 15개에서는 점수 차이는 가장 작아서 0.9점에서 0.5점의 차이를 보였다. 위의 표 4에서 보았듯이 한국인 채점자 그룹의 평균 점수가 원어민 그룹의 평균 점수보다 0.5점이 높았다는 사실에 비추어 볼 때, 위 점수 차이는 실상 유의미하다고 보기 어려울 수도 있다. 이 점을 염두에 두고 토큰들을 살펴보면, 이 부류의 반 이상을 차지하는 경우가 'finish it'에서 'it'를 탈락시키고 'finish'만 쓰거나 연음이 된 것을 한 단어로 잘못 들어서 'finished'로 쓴 경우였는데 한국인 채점자 그룹이 원어민 그룹보다 평균 0.6점 더 높은 점수를 주었다. 또 다른 예로는 'studying'을 두 단어로 잘못 듣고 'study in'으로 쓴 경우였다. 이 부류에 속한 토큰들을 공통적으로 살펴볼 때 한국인 채점자들은 주요한 내용어를 제대로 듣고 쓰는 것을 중요하게 여긴다고 볼 수 있겠다.

3.3. 언어학적 오류 유형에 따른 결과 분석

위에서 보았듯이 25문항에 대한 30명의 수험자의 답안은 총 750개의 답안을 발생시켰는데, 그 중에서 오류가 한 가지라도 있는 오답은 총 263개였다. 그러나 오답에 따라서는 위의 표 2에 제시된 언어학적 오류들의 유형을 보면 알 수 있듯이 그 중에서 두 가지 이상의 오류가 포함되는 경우가 있기 때문에, 많은 오류들이 중복되어 체크될 수 있다. 이렇게 중복을 감안하여 12가지 오류에 해당하는 오답들을 모두 합쳐보면 총 579개가 되었다. 표 2에 제시된

순서대로 오류의 유형에 따른 결과를 논의해 보겠다.

맨 먼저 철자법(Spell) 오류는 총 62건으로서 69번이 11건의 철자법 오류를 보여서 가장 오류 빈도가 높았으며, 그 다음으로는 55번 문항과 56번 문항에서 7건의 철자 오류가 발견되었다. 오류가 발생한 단어로는 'write, address, studying, finish, tomorrow' 등이 있다.

그 다음 유형으로 접미사에 관한 오류를 세 종류로 나누어 분석하였는데 접미사 탈락(Sf_d), 접미사 첨가(Sf_a), 접미사 부조화(Sf_mm)이다. 우선 접미사 탈락은 총 63건으로 54번 문항에서 17건이 발생하여 최다의 경우이고, 그 다음으로 63번 문항에서 16건이 발생하였다. 현재완료형에서 과거분사 'finished'의 *-ed* 접미사가 탈락된 경우가 가장 빈번한 오류였으며, 문법적으로 접미사라고 볼 수는 없지만 'You'd'의 축약된 조동사 *'d*가 탈락된 것도 이 범주에 넣었다. 접미사 첨가는 총 21건으로서 당연히 탈락에 비해 빈도가 현저히 낮게 나타났다. 가장 높은 빈도를 보이는 문항들이 56번의 9건과 58번의 8건으로서 두 문항에서 모두 'finish it'이 연음되어 발음되는 것을 과거형으로 잘못 알아들은 경우로 보인다. 따라서 이 경우에는 합체 오류가 중복되어있다. 그리고 접미사 부조화 오류는 총 3개이며 53번 문항에 있는 'played' 의 과거형 어미 *-ed* 대신 *-s* 또는 *-ing*가 오답으로 쓰인 경우이다.

단어에 대해서도 접미사 오류와 마찬가지로 탈락, 첨가, 부조화 오류를 살펴보았다. 단어 탈락 오류는 총 50건으로 56번 문항에 11건과 65번 문항에 9건이 가장 빈도가 높은 경우였으며 탈락된 단어는 주로 'it'나 'at'와 같은 기능어이다. 단어 첨가 오류는 총 15건으로 접미사의 경우처럼 탈락보다 첨가의 빈도수가 월등하게 낮았다. 단어 첨가 오류가 가장 빈번한 경우는 55번 문항에서 'studying'이 두 단어로 잘못 분석되어 전치사 'in'이 첨가된 결과가 4건 나온 경우와 60번 문항에서 'have rained'가 연음되어 발음되는 과정에서 부정관사 'a'가 있는 것처럼 잘못 듣고 답을 쓴 경우 3건이 있었다. 단어 부조화 오류는 총 121개로서 63번 문항에서 20건이 가장 많았고 그 다음으로는 71번 문항에서 17건 발생했다. 단어 부조화의 빈번한 예는 63번에서'too'를 'to'로 바꾸어 쓰거나 71번에서는 'won't'를 'will'로 교체하는 오류였다.

다음으로 위의 단어 오류의 경우들을 기능어(function word)와 내용어(content word)의 경우로 구분해서 빈도를 살펴보았다. 단어 오류들 중에서 기능어에 속하는 단어들은 총 123건이었으며 내용어에 속하는 단어들은 총 64건이었다. 기능어 오류는 63번, 62번 문항에서 가장 많은 오류가 발생했으며 '(you)'d, could, 'll, did, do, can't, can, would'와 같은 조동사에 대한 오류였다. 그리고 내용어 오류는 63번에서 16건과 69번에서 13건이 최다 경우이다. 대표적인 단어들은 'stop, eating, too, write, address, there' 등이다.

그 다음에는 71번부터 75번까지 문항들에 포함된 조동사의 부정 형태가 반대로 긍정 형태로 나타난 오답이 경우들을 보았는데 총 28건으로서 71번의 'won't'가 'will'로 17건이나 나타났고 72번의 'can't'도 'can'으로 쓰인 경우가 5건 있었다. 반대로 65번 문항의 긍정 조동사 'can'을 'can't'로 잘못 쓴 경우도 2건 있었다.

마지막으로 합체(coalescence) 오류와 분리(division) 오류를 살펴보았는데 합체 오류는 두 단어를 한 단어로 잘못 들어서 생기는 오류로서 주로 연음을 제대로 청취하지 못한 결과로 생기며 중대한 소통상의 오류의 원인이 된다. 총 21건의 합체 오류가 발견되었는데 56번, 58번 문항에서 'finish it'을 'finished'로 쓴 경우가 각각 5건씩 있었고, 그 다음으로는 60번 문항의 'have rained'를 'brained, bringed' 등의 의미가 통하지 않는 단어로 합체한 경우가 4건 있었고, 55번 문항에서는 'have been'을 'having'으로 합체한 경우가 3건 있었다. 끝으로 분리 오류의 경우는 총 8건으로서 축약된 조동사를 도로 분리해서 씀으로써 문법적으로는 아무런 하자가 없는 결과를 가져오는 경우(58번의 'I'll'을 'I will' 로 쓴 경우)가 5건으로 가장 많았지만, 55번 문항의 'studying'을 'study in'으로 잘못 분리하여 씀으로써 문법적으로 맞지 않는 문장을 도출하게 되는 결과도 2건 있었다.

3.4. 언어학적 오류와 채점방식별 평가결과의 상관관계

앞의 3.2에서는 채점방식들 간에 크게 점수가 벌어지는 데이터들을 모아서 어떤 요인들이 그와 같은 차이를 가져왔는지 분석해 보았다. 이 장에서는 이 연구에서 다루고 있는 12가지 언어학적 오류의 유형과 세 가지 채점방식 간의 점수 차이의 연관성을 통계적으로 살펴보기로 하겠다. 각각의 오류유형에 해당하는 데이터들을 모아서 각 채점방식별로 점수의 평균값과 표준편차를 내고 결과로 나온 평균값의 차이가 통계적으로 유의한 차이를 보이는가를 분석하여 표 5에 제시하였다. 세 가지 채점방식 차이를 볼 때는 Friedman 분석을 시행했으며, 사후분석으로는 Bonferroni 방법을 사용하였고 보정된 유의수준은 0.017로 나타났다.

우선 표 5를 해석하는 방법을 철자법 오류의 예를 들어 설명해 보겠다. 철자법 오류가 있는 총 62개의 토큰에 대한 채점방식 간 점수를 냈더니 6점 만점일 때 원어민 채점자들의 평균 점수는 4.83점, 한국인 채점자들의 평균 점수는 4.76점, 그리고 전산점수는 3.94점으로 나왔다는 뜻이며, 괄호 안의 숫자는 표준편차를 의미한다. 즉, 철자법 오류의 경우 채점방식별 점수에 유의한 차이를 보였고($p<.05$), 원어민, 한국인, 전산 점수의 순으로 높은 점수를 주었음을 알 수 있다. 그러나 사후분석 결과에 의하면 원어민과 한국인 점수에는 유의한 차이가 없는 것으로 나타났고($p=.029$), 원어민과 전산 ($p=.000$), 한국인과 전산 ($p=.000$) 사이에는 유의한 차이가 있는 것으로 나타났다. 즉, 수기 채점에서는 철자법 오류가 전산 채점에서의 경우에 비해 감점이 작은 것으로 드러났다.

표 5. 오류별 채점방식 간의 점수 차이 M(SD)

오류유형	원어민	한국인	전산	χ^2	p
Spell	4.83(0.78)	4.76(0.73)	3.94(0.90)	81.139	.000
Sf_d	4.45(0.64)	4.37(0.56)	3.72(0.70)	66.124	.000

Sf_a	3.92(0.57)	4.10(0.50)	3.39(0.84)	22.988	.000
Sf_mm	4.53(1.24)	4.67(1.10)	4.05(0.54)	1.273	.529
Wd_d	4.09(0.78)	4.04(0.78)	3.29(0.91)	45.672	.000
Wd_a	4.15(0.71)	4.06(0.62)	3.80(0.92)	3.763	.152
Wd_mm	4.42(0.70)	4.15(0.68)	3.77(0.84)	80.630	.000
Func	4.40(0.70)	4.26(0.66)	3.79(0.89)	66.895	.000
Cont	4.24(0.75)	3.89(0.66)	3.36(0.81)	68.167	.000
NP	4.50(0.69)	4.11(0.49)	4.13(0.66)	17.236	.000
Coal	3.76(0.65)	3.83(0.66)	3.02(0.87)	24.025	.000
Div	4.73(1.37)	4.91(1.18)	4.53(1.58)	8.000	.018

$p<.05$

접미사 오류에서는 우선 접미사 탈락의 경우(n=63)에도 채점방식별 점수에 유의한 차이를 보였는데, 원어민, 한국인, 전산 점수 순으로 높은 점수가 나왔으며, 원어민과 한국인 점수에도 유의한 차이가 있는 것으로 나타났다(p=.014). 반면에 정답에는 없는 접미사가 첨가(n=21)되었을 때에는, 세 가지 채점방식 간의 점수 차이는 한국인, 원어민, 전산 순으로 높게 나타났다. 한국인 채점자가 원어민 채점자보다 높은 점수를 주었으며 이 차이도 유의한 차이로 나타났다(p=.037). 그러나 접미사 부조화 오류는 총 토큰의 수가 너무 작은(n=3) 탓인지 채점방식 간에 아무런 유의한 차이를 보이지 않았다.

단어 오류 중에서 단어 탈락 오류(n=50)의 경우에는 원어민, 한국인, 전산 점수 순으로 높은 점수가 주어졌고 채점방식 간의 점수에 유의한 차이를 보였으나, 원어민(4.09)과 한국인(4.04) 점수 사이에는 유의한 차이가 없는 것으로 나타났다(p=.442). 단어 첨가 오류의 원어민, 한국인, 전산 점수 순으로 높은 점수가 나왔지만, 데이터가 빈약한(n=15) 때문인지 통계적인 유의한 차이는 나타나지 않았다(원어민-한국인 p=.704; 원어민-전산 p=.035; 한국인-전산 p=.078). 단어 부조화 오류(n=121)에서는 원어민, 한국인, 전산 점수 순으로 높은 점수가 주어졌으며 이 차이는 통계적으로 유의한 차이를 보였고, 모든 경우에 유의한 차이가 있는 것으로 나타났다(p=.000).

해당 단어가 기능어인가 내용어인가에 따라 오류를 분류한 경우에는 두 경우 모두에서 원어민, 한국인, 전산 점수 순으로 높은 점수를 주었으며 그 순서는 유의한 차이를 보였다. 원어민과 한국인 점수 사이에서도 기능어 오류(p=.001)에서나 내용어 오류(p=.000) 모두에서 유의한 차이가 있는 것으로 나타났다.

다음으로 의미적 오류인 긍정과 부정이 교체된 오류(n=28)에서도 원어민, 한국인, 전산 점수 순으로 높은 점수가 나타났고 그 차이는 유의한 것으로 분석되었다. 원어민과 한국인의 점수 차이 역시 유의한 차이가 있는 것으로 나타났다(p=.000). 뒤에서 살펴볼 오류별 점수 순위에 있어서도 원어민의 경우는 한국인의 경우보다 이 의미적 오류의 점수 순위가 상대적으로 높았다.

마지막으로 합체오류(n=21)와 분리오류(n=8)에서는 동일한 결과가 나타났는데, 세 가지 채점방식 사이에서는 한국인, 원어민, 전산 점수 순으로 높은 채점 결과를 보였으며 이 결과는 유의한 차이로 분석되었으나, 한국인과 원어민 간의 점수 차이는 두 가지 경우에서 모두 유의한 차이가 없는 것으로 나타났다(p=.246; 분리오류 p=.068).

표 5의 결과를 시각적인 비교가 보다 용이하도록 선도표로 나타낸 결과가 아래의 그림 2와 같다.

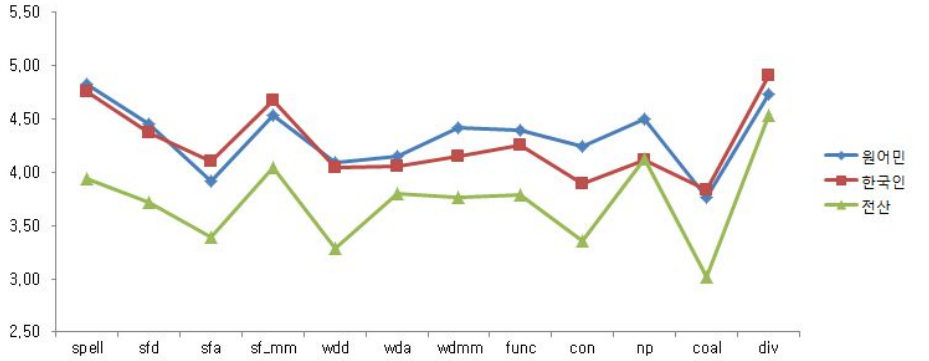


그림 2. 오류별 채점방식 간의 점수 비교

전산채점 방식에 의한 점수는 수기채점 방식에 비해 오류의 유형에 관계없이 낮은 점수가 나온 것을 알 수 있다. 수기채점 방식의 두 가지 점수를 비교해 보면 대부분의 오류에서는(12개 유형 중 10개) 원어민의 점수가 한국인의 점수보다 높거나 거의 비슷한 수준으로 나타났다. 단 2개(접미사 첨가오류와 부조화오류)에서만 한국인의 점수가 높게 나타났는데, 그 중에서 접미사 부조화오류는 단 3개의 토큰만이 있었으므로 그 차이의 통계적 유의미성을 검증할 수 없는 상황이었다. 접미사 첨가오류의 경우는 총 21건으로서 매우 예외적으로 한국인들이 원어민들보다 평균적으로 높은 점수를 결과가 나온 셈이다. 정확한 원인은 규명할 수 없으나 전체적인 상황을 볼 때 한국인 채점자들은 문법적으로는 틀리지만 사소한 접사가 첨가되어 내용에 큰 영향을 미치지 않을 때 원어민 채점자들에 비해 감점을 덜 한 것으로 보인다.

지금까지 언어학적 오류에 따른 채점방식 간의 평가결과 차이에 연관성 여부를 살펴보았다. 분석 결과 분명히 세 가지 채점방식 간에 유의한 점수 차이가 존재한다는 것은 검증되었다. 이와 같은 결과를 기반으로, 오류에 따른 상대적인 점수 차이와 그 원인을 분석해 보기 위해 채점방식별로 12가지 오류에 대한 (평균)점수의 순위를 매겨보았는데 그 결과는 아래의 표 6에 제시하였다.

오류별로 채점방식 간에 점수 차이를 살펴보았을 때는 유의한 점수 차이가 있었으나, 표

6에서 보듯이 오류별 점수의 순위를 따져보았을 때는 세 가지 채점방식 간에 비교적 유사한 순위를 보인다는 사실이 드러났다. 문법적으로 거의 오류라고 볼 수 없는 경우가 포함되어있는 분리(Div)오류가 세 가지 방식 모두에서 가장 높은 순위인 1, 2위를 보이고 있으며, 연음에 대한 잘못된 청취와 그 부분의 문법/어휘가 제대로 습득되지 않아서 일어날 수 있는 합체(Coal)오류는 세 가지 방식 전체에서 가장 감점을 많이 받아서 최하위 순위를 기록하고 있다.

표 6. 채점방식별 오류에 대한 점수 순위

점수 순위	원어민	한국인	전산
1	*Spell*	Div	Div
2	Div	*Spell*	**N-P**
3	Sf_mm	Sf_mm	Sf_mm
4	**N-P**	Sf_d	*Spell*
5	Sf_d	Func	Wd_a
6	Wd_mm	Wd_mm	Func
7	Func	**N-P**	Wd_mm
8	Cont	Wd_a	Sf_d
9	Wd-a	Wd_d	Sf_a
10	Wd_d	Sf_a	Cont
11	Sf_a	Cont	Wd-d
12	Coal	Coal	Coal

오류가 발생한 형태소가 접사이냐 단어이냐에 따라 점수 차이가 발생할 것이라는 가정하에 이 경우를 비교하여 보면, 탈락이나 부조화의 오류에서는 세 가지 방식 모두에서 접미사의 경우가 단어의 경우보다 점수 순위가 높았다. 즉, 정답에 있는 접미사가 탈락된 오류보다 정답에 있는 단어가 탈락된 오류가 더 심각한 오류로 판단되었다고 해석할 수 있을 것이다. 이는 접미사가 가지고 있는 정보가 단어의 정보보다는 상대적으로 덜 중요하기 때문에 나타난 당연한 결과라고 볼 수 있겠고, 채점방식의 신빙성을 더해주는 요인이 될 것이다. 반면에 첨가(addition)오류의 경우에는 정반대의 결과가 세 방식 모두에서 나타났다. 즉, 정답에 없는 단어가 첨가되었을 때보다 정답에 없는 접미사가 첨가된 오류가 더 심각한 오류로 판단되었다는 것이다. 실제 오류 토큰들을 살펴보니 단어 첨가 오류에서는 15건 중 두 건을 제외하고는 모두 ‘in, to, a’등의 전치사나 부정관사와 같은 기능어가 첨가된 것을 알 수 있었고, 접미사 첨가 오류 21건 중 한 건을 제외하고는 모두 조동사 뒤에 오는 원형동사에 과거 접미사 ‘-ed’를 첨가한 오류들이었다.

다음으로 단어 오류 중에서 오류가 발생한 단어가 관사, 전치사, 조동사, 인칭 대명사 등의 기능어에 속하는 단어인지 아니면 일반명사, 일반동사, 형용사, 부사 등의 내용어에 속하는 단어인지를 구분하여 분석해 본 기능어(Func) 오류 대 내용어(Cont) 오류를 비교해 보았다. 이 변수에 대한 결과는 채점방식 간에 상하위 여부는 동일하여 내용어 오류가 기능어 오류보다 하위로 나타났으나 세부 사정은 각각 다르게 나타났다. 내용어가 기능어에 비해 중요도가 높은 것은 당연하기에 두 오류 사이의 순위는 예상과 일치했다. 그러나 한국인 채점자의 기능어 오류가 5위이고 내용어 오류가 11위를 차지하여 상당한 순위 차이와 0.37점의 평균 점수 차이를 보이는 반면에 원어민 채점자의 경우는 7, 8위로서 미세한 순위 차이와 0.16점의 점수 차이만을 보이는 대조를 이루었다. 한편, 기능어와 내용어의 구분이 반영되지 않는 채점방식을 채택하는 전산채점방식에서는 기능어 오류가 6위, 내용어 오류가 10위를 차지하였고, 평균 점수 차이 또한 0.43점으로 가장 큰 차이를 보였다. 두 가지 수기채점방식을 비교 분석해 볼 때, 한국인 채점자들이 원어민 채점자들보다 단어의 중요도에 더 큰 비중을 둔다고 볼 수 있겠으나, 전산채점의 결과에 대한 해석은 더 많은 자료에 대한 연구가 필요할 것으로 보인다.

채점방식 간에 상이한 순위를 보이는 오류가 두 가지 있는데, 철자법 오류와 의미적 오류인 부정-긍정 교체 오류가 그것들이다. 먼저 부정-긍정 교체 오류를 살펴보면, 원어민 채점방식에서는 4위, 한국인에서는 7위, 그리고 전산에서는 2위의 점수 순위를 보였다. 원어민 채점자들은 철자법 오류, 분리 오류, 접미사 부조화 오류 다음으로 부정-긍정 교체 오류 점수를 높이 준 반면에 한국인 채점자들의 경우에는 부정-긍정 교체 오류가 접미사 탈락 오류나 단어 부조화 오류보다도 더 낮은 점수 순위에 놓였다. 즉, 한국인 채점자들은 정답의 의미를 정반대로 만드는 오류를 원어민들에 비해 훨씬 더 심각한 오류로 받아들인다고 해석할 수 있겠다. 한편 전산채점의 경우에는 부정-긍정 교체 오류가 분리 오류 바로 아래인 2위를 차지하여 전산채점으로는 의미상의 오류를 제대로 평가할 수 없다는 결론을 내릴 수 있다.

마지막으로, 철자법 오류를 살펴보면, 수기채점의 경우 우선 원어민에서는 가장 높은 점수 순위를 보였고 한국인에서도 무오류에 가까운 분리 오류 다음인 2위를 차지해서, 철자법 오류는 받아쓰기에 대한 수기채점에서 가장 사소한 오류로 평가됨을 알 수 있다. 반면에 전산 채점방식에서는 사소한 철자법 오류를 분리해서 평가하는 공식이 없으므로 부정-긍정 교체 오류보다도 순위가 낮은 4위에 해당하는 결과가 드러났으므로, 향후 전산 채점방식을 개정할 때 염두에 두어야 할 사안이라고 볼 수 있다.

4. 결론

본 연구에서는 영어 받아쓰기 평가를 위한 전산채점의 신뢰성을 파악하기 위해서 전산채점의 결과를 수기채점의 결과와 비교 분석하여 뚜렷한 차이들을 찾아내고 그 차이를 만드는

요인들을 밝히고자 하였다. 국내에서 외국어를 가르치는 교사가 영어가 모국어가 아닌 한국인과 영어가 모국어인 외국인이 거의 대부분이기 때문에 이 두 그룹의 수기채점과 전산채점을 비교하였다.

연구 결과를 요약하면 전산채점에서는 작은 단어나 중요한 단어를 구별 못 하기 때문에 작은 단어 오류에 대한 과중한 감점이 됨을 알 수 있었다. 하지만 주요 단어의 오류가 있거나 탈락이 될 경우 수기채점자들은 감점을 많이 했으나 전산채점에서는 단어 한 개의 실수로 상대적으로 덜 중요한 기능어의 오류와 동일하게 처리되는 문제가 부각되었다. 원어민은 비슷한 발음의 다른 단어를 쓴 경우 상대적으로 높은 점수를 준 반면에 한국인은 이러한 오답을 심각한 오류로 여기는 경향이 있었다. 따라서 철자법 오류에 대해서도 원어민의 점수가 후하였고, 전산채점은 가장 엄격하게 점수를 주었다. 한국인 채점자들은 주요한 내용어를 제대로 듣고 쓰는 것을 중요하게 여긴다고 볼 수 있었다. 즉 문맥을 제대로 이해하면 점수를 후히 주고, 작은 실수라도 문장의 의미를 왜곡 시키면 과중한 감점을 함을 알 수 있었다. 따라서 기능어와 내용어의 오류에서 한국인의 채점이 가장 엄격하였다. 반면에 접미사 오류에서는 한국인의 점수가 원어민의 점수보다 후하였다.

전체적으로는 원어민 수기 채점자는 발음적으로 가까우면 약간 틀린 것에 후하게 점수를 준 반면, 한국인 수기 채점자는 스펠링 차이가 별로 없더라도 의미가 확 달라지거나 제대로 듣지 못 했다고 판단하는 경우에는 엄한 점수를 주었다. 반면에 기계식 전산채점은 주요 단어이든 기능어이든, 철자법 오류이든 관계없이 일률적인 점수를 주어 수기 채점자와 크게 벗어나는 결과를 보였다. 또한 수기채점은 전산채점에 대비하여 약간의 비일관성도 보였다. 이러한 비일관성을 벗어날 수 있는 것이 전산화된 채점이기 때문에 수기채점의 장점을 살릴 수 있는 더욱 좋은 알고리즘을 개발할 필요가 있을 것이다.

따라서 앞으로 개발할 알고리즘은 첫째로 철자법 및 접미사 오류에 대해서 오류의 예를 풍부하게 확보하고 이에 대한 점수 배정을 미리 해 놓는 것이 필요할 것이다. 둘째로 기능어와 내용어의 배점을 달리하여 그 오류들이 일괄적으로 계산되지 않도록 하며, 셋째로 긍정과 부정이 교차된 예도 오류의 예를 미리 등록하여 한국인 수기 채점자와 원어민 수기 채점자의 중간 점수가 나올 수 있도록 배점을 해야 할 것이다. 마치 기계번역이 단순히 좋은 알고리즘만이 아닌 얼마나 많은 번역 기억용량(translation memory)이 있는가에 의해서 결과물이 큰 차이를 보이듯이, 많은 테스트를 통하여 오류 데이터베이스를 충분히 확보하는 것이 보다 바람직한 전산채점을 향한 중요한 단계가 될 것으로 보인다.

나아가 단순히 받아쓰기 평가를 위한 채점만 하는 것이 아니라 해당 학습자가 어떤 오류를 보이는 가까지 정확히 짚어 줄 수 있다면, 받아쓰기-채점-후행학습 제안까지 순간적으로 이루어질 수 있어 학습자들이 자신의 약점을 파악하고 그 약점을 중심으로 연습을 하여 빠른 진보를 보일 수 있도록 할 수 있을 것이므로 추후 이 분야에 대해서도 많은 연구가 이루어져야 할 것이다.

참고문헌

김보라, 이규민. (2012). 일반화가능도 이론을 적용한 초등학교 쓰기 수행평가의 총체적 채점과 분석적 채점 방식 비교. *교육학연구, 50*(4), 49-76.

김영철. (1988). 중학생들의 받아쓰기 평가절차에 대한 연구. *Studies in English Education, 3*(2), 147-172.

김원명. (1984). *영어 청취 오류분석*. 한남대학교 대학원 석사학위 논문.

에듀조선영어연구소. (2010). *영어 받아쓰기 공식문제 Workbook. DAT(Dictation Assessment Test. Level 5.* 경기도 성남시: 에듀조선.

이선주. (2006). *멀티미디어를 활용한 받아쓰기 평가 및 채점방식 타당성 분석*. 고려대학교 대학원 석사논문.

이현구, 윤병남. (2007). 중학생 영어학습자의 영어 받아쓰기에서 나타난 음운론적 오류분석 연구. *언어학, 15*(4), 123-146.

정소라. (2008). *점층식 받아쓰기 채점방식의 실용도와 신뢰도*. 고려대학교 교육대학원 석사논문.

최윤정. (2006). *에세이 채점을 위한 컴퓨터 프로그램의 비교분석*. 이화여자대학교 석사논문.

최인철, 임해창, 박정. (2003). 영작문의 전산언어학적 채점 타당성. *멀티미디어 언어교육, 6*(2), 221-241.

최인철, 엄연희. (2001). 작문수행평가에 대한 전문가와 전산언어학적 채점결과의 비교 분석. *멀티미디어 언어교육, 4*(1), 165-184.

Buck, G. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

Brown, H. D. (1980). *Principles of language learning and teaching*. Englewood Cliffs, N.J.: Prentice-Hall.

Duskova, L. (1969). On sources of error in language learning. *IRAL*, 7, 35-79.

James, C. (1977). Judgments of error gravities. *English Language Teaching Journal, 31*(2), 116-124.

McNamara, T. F., & Candlin, C. N. (1996). *Measuring second language performance*. London: Longman.

Oller, J. W. Jr. (1972). *Dictation as a test of ESL proficiency*. New York: McGraw Hill

Oller, J. W. Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Paulston, C. B., & Bruder, M. N. (1976). *Teaching English as a second language: Techniques and procedures*. Cambridge, Mass.: Winthrop Publishers, Inc.

Postovsky, V. A. (1974). Effects of delay in oral practice at the beginning of second language learning. *The Modern Language Journal, 58*, 229-239.

Rivers, W., & Templerly, S. (1978). *A practical guide to the teaching of English as a second language*. New York: Oxford University Press.

이보림
570-749 전북 익산시 익산대로 460
원광대학교 인문대학 영어영문학과
전화: (063)850-6875
이메일: brlee@wku.ac.kr

최종헌
463-742 경기도 성남시 분당구 새마을로 257번지
성남영어마을 원장
전화: (031)725-5633
이메일: joshuachoi@educhosun.co.kr

Received on September 11, 2014
Revised version received on December 1, 2014
Accepted on December 10, 2014