Korean Journal of English Language and Linguistics, **Vol 25, August 2025, pp. 1048-1065**
DOI: 10.15738/kjell.25..202508.1048

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

한국인 학습자 영어 말하기 자동 평가를 위한 문법 다양성 척도 연구*

Min-Chang Sung (Gyeongin National University of Education) · **Jin-Hwa Lee** (Chung-Ang University) · **Heyoung Kim** (Chung-Ang University) · **YunDeok Choi** (Chungnam National University)

Received: June 10, 2025
Revised: July 23, 2025
Accepted: August 5, 2025

Sung, Min-Chang (First author)
Associate Professor, Department of English Education, Gyeongin National University of Education
Email: mcsung@ginue.ac.kr

Lee, Jin-Hwa (Corresponding author)
Professor, Department of English Education, Chung-Ang University
Email: jinhlee@cau.ac.kr

Kim, Heyoung (Co-author)
Professor, Department of English Education, Chung-Ang University
Email: englishnet@cau.ac.kr

Choi, YunDeok (Co-author)
Assistant Professor, Department of English Education, Chungnam National University
Email: yundeokchoi@cnu.ac.kr

*This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5A2A03086569).

ABSTRACT

Sung, Min-Chang, Jin-Hwa Lee, Heyoung Kim and YunDeok Choi. 2025. A study on grammatical diversity measures for automated English speaking assessment of Korean learners. *Korean Journal of English Language and Linguistics* 25, 1048-1065.

This study investigates various measures of grammatical diversity for automated speaking assessment (ASA) with Korean learners of English. Data were extracted from the Korean monologue set in the International Corpus of Asian Learners of English and classified into two proficiency groups. Using two NLP toolkits, i.e., the Biber tagger and the argument structure construction annotator, we measured the grammatical diversity of the speaking data based on six features over three levels (i.e., word, phrase, and clause), and conducted correlation and binomial logistic regression analyses. The results reveal significant correlations among the features, particularly between those related to part-of-speech and the others. It is also found that only the clause-level feature of the subordination type frequency significantly predicts proficiency levels. These findings provide insights into the potential of grammatical diversity as a valuable metric in ASA systems for Korean learners of English.

KEYWORDS

automated speaking assessment, grammatical diversity, English proficiency, L2 corpus, NLP-based analyses

1. 서론

자연어 처리 기술의 발달과 인공지능을 활용한 대규모 언어 모델(LLM: Large Language Model)의 출현 등으로 영어 학습자의 자동 말하기 평가(ASA: automatic speaking assessment)가 전 세계적으로 보편화되고 있다. 영어 자동 말하기 평가는 영어 말하기 문항을 제시하고, 영어 학습자의 음성 발화 데이터를 수집 및 전사하여, 주요 말하기 능숙도 관련 지표들을 분석 및 정량화한 뒤, 평가 결과와 피드백을 산출하는 일련의 과정을 포함한다(Xu et al. 2021, Zechner and Evanini 2019). TOEFL과 DET(Duolingo English Test) 등 국제적인 대규모 영어 능숙도 시험뿐만 아니라, 영어 학습용 챗봇과 온라인 영어 학습 플랫폼 등 영어교육 콘텐츠에서도 영어 말하기 자동 평가를 도입하는 추세이다(남보라 외 2024, 이진화 외 2023).

영어 말하기 자동 평가를 위해서는 유창성, 문법 정확성, 어휘 사용 등 평가 영역별로 분석 자질과 척도를 결정하고, 정량화된 데이터에 입각하여 학습자 말하기의 수준을 평가할 수 있어야 한다. 예를 들어, 유창성(Fluency) 영역에서는 '흐름'이나 '속도' 등의 자질을 평가할 수 있는데, 이를 위해 '발화의 연속 및 단절'이나 '분당 단어 수(WPM: Words Per Minute)'와 같은 양적 척도를 활용할 수 있다(이진화 외 2023, Xu el al. 2021). 어휘(Vocabulary) 영역에서는 '어휘 다양성'과 '어휘 정교성(sophistication)' 등의 자질을 '단어 유형 수'와 '참조 코퍼스 내 어휘 빈도' 등의 척도로 평가할 수 있다. 이와 같이 영어 말하기 자동 평가의 주요 영역에 대한 분석 자질과 척도는 상당한 수준으로 체계화되어 있다.

그러나 일부 평가 영역에서는 분석 자질과 척도에 대한 합의가 미진하며, 분석 자질과 척도가 체계적인 이론적 논의에 바탕을 두지 않은 채 자연어 처리(NLP) 기술의 용이성과 임의적인 통계 분석에서 관측되는 설명력에 의존하는 경우가 있다. 본 연구는 그 대표적인 사례로 '문법 다양성(Grammatical Diversity)' 영역을 탐구하고자 한다. 문법 다양성은 문법 복잡성의 하위 개념으로서 정의되어 왔는데, 다양한 문장 패턴(Skehan 1996)이나 발화를 구성하는 형태의 범위(Ortega 2003) 등으로 기술된다.

영어 말하기 평가에서 문법 다양성은 학습자의 능숙도와 문법 패턴의 다양성이 비례한다는 현상에 근거하고 있다(ACTFL 2012, 2024, Council of Europe 2020). 즉, 낮은 수준의 학습자는 소수의 문법 패턴을 반복적으로 활용하는 반면, 높은 수준의 학습자는 다양한 문법 패턴을 적재적소에서 활용한다(Sung and Kim 2022). 문법 다양성은 발화 단위에 따라 단어 수준, 구 수준, 절 수준 등에서 논의되어 왔다. 예를 들어, 단어 수준에서는 다양한 단어 형태(예: 동사 과거형, 현재형 등)와 품사(POS: Part of Speech)를 사용하는지를, 구 수준에서는 다양한 구조를 사용하는지를, 절 수준에서는 종속절 유형이 다양한지를 평가할 수 있다. 이와 관련하여, 실제 영어 자동 말하기 평가에서는 품사 활용 양상(예: POS bigrams), 종속절(dependent clause)의 분포 등을 문법 다양성의 평가 척도로 활용하고 있다(Bhat and Yoon 2015, Chen et al. 2018).

문법 다양성 지표들은 영어 말하기 능숙도를 평가하는 데 유용하게 활용되고 있으나, 이론적 논의와 실용성 검토를 통해 보완될 필요가 있다. 우선, 단어, 구, 절 수준에서 문법 다양성이 측정될 수 있음에도 불구하고, 특정 수준에서만 문법 다양성을 측정하는 현상을 비판적으로 고찰할 필요가 있다(Zechner and Evanini 2019). 나아가 각 수준에서 하나의 측정 지표를 임의로 설정하기보다는 영어 학습자의 말하기에 대한 종합적인 이론적 고찰에 입각하여 복수의 지표를

탐색할 필요가 있다. 예를 들어, 품사의 배열이 문법 구조로 정의될 수 있는지, 종속절과 주절 중 무엇이 절 수준의 문법 다양성을 대표하는지, 그리고 문법 다양성 지표들이 어떠한 유사성과 차별점을 보이는지 등에 대한 논의가 필요하다. 그러나 한국인 영어 말하기의 문법 평가는 여전히 다양성보다는 정확성에 치중되어 있고(신동광 외 2015), 한국인 영어 말하기의 문법 다양성 연구 또한 문장 구조와 같은 특정 문법 항목에 제한되어 있는 상황이다(Sung and Kim 2022).

본 연구는 이러한 문제의식에 근거하여 한국인 영어 말하기 자동 평가에 적합한 문법 다양성 척도를 탐구하고자 한다. 이를 위해, 두 개의 능숙도 집단에서 수집된 영어 말하기 데이터의 문법 다양성을 단어, 구, 절 수준에서 각각 2개의 다른 지표로 분석하여, 지표 간 상관관계와 각 지표의 능숙도에 대한 설명력을 분석한다. 구체적인 연구 질문은 다음과 같다.

1) 한국인 영어 말하기 데이터에서 단어, 구, 절 수준의 문법 다양성 지표들은 유의미한 상관관계를 보이는가?

2) 한국인 영어 말하기 데이터에서 단어, 구, 절 수준의 문법 다양성 지표들은 영어 능숙도를 어느 정도로 설명하는가?

2. 문법 다양성과 영어 말하기 자동 평가

문법 다양성은 문법 복잡성(Grammatical Complexity)의 하위 개념으로, 문법 정교성(Grammatical Elaboration)과 비견된다. 문법 다양성은 언어 산출에서 나타나는 문법 형태의 폭넓은 사용과 관련이 있고, 문법 정교성은 각 형태가 얼마나 정교하고 복잡한 구조 안에서 활용되는지와 관련이 있다(Kyle 2016, Ortega 2003). 예를 들어, 종속절에서 문법 다양성이 높은 학습자는 다양한 형태 및 의미 유형의 종속절을 사용하지만, 문법 정교성이 높은 학습자는 각 종속절의 길이가 길고 구조가 복잡할 것이다.

문법 다양성은 영어 습득에서 일정한 발달 패턴을 보이는 것으로 알려져 있다(Dulay and Burt 1974, Krashen and Terrell 1983). 우선, 영어 원어민에 비해 영어 학습자들은 제한된 범위의 문법 형태를 사용하는 양상을 보인다. 또한 영어 능숙도와 문법 다양성 간의 상관 관계가 관측되었는데, 원어민 아동보다 원어민 성인이, 낮은 수준의 영어 학습자보다 높은 수준의 영어 학습자가 더 다양한 문법 형태를 사용하는 경향이 있다(Ellis and Ferreira-Junior 2009, Kyle and Crossley 2017, Lee and Kim 2011). 이와 같은 능숙도와 문법 다양성 간의 정적 상관관계는 한국인 영어 학습자의 발화에서도 관측되었다(Park and Sung 2024, Sung and Kim 2022). 예를 들어, 낮은 수준의 영어 학습자는 타동사의 목적어에 명사구만을 활용하는 반면, 높은 수준의 영어 학습자는 타동사의 목적어에 동명사구와 명사절 등 보다 다양한 문법 형태를 활용한다(Choi and Sung 2020).

언어 습득에서 밝혀진 문법 다양성 발달 양상은 제2언어 능숙도를 기술하는 이론 체계에도 적용되고 있다. 대표적으로 유럽공통참조기준(CEFR: Common European Framework of Reference for Languages)은 언어 능숙도를 여섯 등급(A1, A2, B1, B2, C1, C2)으로 구분하고, 각 등급의 '언어 능력(Linguistic competence)'을 기술할 때 '일반 언어 범위(General linguistic range)'라는 범주로 문법 다양성을 적용하고 있다. 이 중 가장 낮은 등급인 A1과 가장 높은 등급인 C2의 문법 다양성은 다음과 같이 기술되어 있다(Council of Europe 2020, pp. 130-131).

A1: Has a very basic range of simple expressions about personal details and needs of a concrete type.

C2: Can exploit a comprehensive and reliable mastery of a very wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity. No signs of having to restrict what they want to say.

즉, 능숙도가 낮은 영어 학습자는 매우 기본적인 범위의 간단한 문법 형태를 활용하는 반면, 능숙도가 높은 영어 학습자는 광범위한 언어 표현을 일관성 있게 활용한다.

이와 같은 능숙도 구분 체계에 입각하여, 영어 말하기 시험에서도 학습자의 문법 다양성을 평가하고 있는데, 시험의 목적과 특징에 따라 다양한 개념과 방법으로 문법 다양성을 측정 및 평가하고 있다. 예를 들어, TEPS Speaking은 Part 3 문항의 평가 항목 중 하나인 정확성(Accuracy)에서 '구문의 적절한 사용'이라는 세부 평가 항목으로 문법 다양성을 채점하는 반면, IELTS는 모든 문항에서 문법 다양성을 평가하여 '문법 범위와 정확성(Grammatical range and accuracy)' 영역의 10분위(0~9) 중 하나의 등급을 배정하는 데 활용한다.

최근에 개발된 자동 영어 말하기 평가 시스템에서도 문법 다양성에 대한 정보를 활용하기 위하여 다양한 분석 체계와 측정 기술이 활용되고 있다. 예를 들어, Chen et al.(2018)의 연구는 영어 말하기 자동 평가 모델인 SpeechRater ver. 5에 활용될 평가 시스템을 개발하는 과정에서 10여 개의 문법 지표들이 능숙도 판정에 가지는 기여도를 분석하였다. 그 결과, 2개의 문법 지표를 최종 평가 모델에 6.3% 비중으로 포함하였는데, 대부분의 비중이 문법 다양성에 해당하는 품사 이진(POS bigram) 유형의 빈도 분포에 배정되었다. 여기서 품사 이진이란 연속된 두 단어의 품사 태그 쌍을 의미하는데, 예를 들어 "The_DT cat_NN sleeps_VBZ" 문장의 품사 이진은 <DT, NN>와 <NN, VBZ>이다. Zechner와 Evanini(2019) 또한 다양한 문법 지표 중 품사 배열 패턴이 인간 채점자가 매긴 TOEFL iBT 말하기 능숙도 종합 점수와 가장 높은 상관관계를 보였음을 밝혔다.

그 결과, 실제 여러 유형의 말하기 자동 평가 시스템에서 품사 배열 분포를 문법 다양성의 주요 지표로 활용하는 추세이다(Bhat et al. 2014, Chen et al. 2018, Khristoforov et al. 2020, Yoon and Bhat 2012). 그러나 품사 배열 분포에 대한 의존은 한국인 영어 학습자의 말하기 자동 평가 시스템 구축과 관련하여 여러 가지 시사점을 가진다. 우선 문법 다양성은 단어, 구, 절, 문장 등 여러 층위에서 논의될 수 있는데, 품사 배열은 주로 구 단위의 분석에 제한된 것으로 보인다. 또한 품사 배열이 제2 언어 습득이나 말하기 수행에서 의미하는 바가 명확하지 않기 때문에, 상위 수준의 학습자와 유사한, 혹은 유사하지 않은 품사 배열 분포를 보였다는 것을 평가 응시자에게 명확히 설명하고 적절한 피드백을 제공하기 어렵다(Zechner and Evanini 2019). 나아가 한국인 학습자 간의 말하기 능숙도를 분별하는 데 품사 배열 분포가 유의미한 지표로 활용될 수 있는지에 대한 검증 또한 필요하다. 그러나 한국인 영어 학습자 대상의 말하기 자동 평가는 여전히 문법 다양성보다는 문법 정확성이나 문법 정교성에 치중되어 있는 상황이다(신동광 외 2015).

따라서 본 연구는 한국인 학습자의 영어 말하기 데이터를 단어, 구, 절 수준의 문법 다양성으로 종합 분석하여, 지표 간의 상관관계와 능숙도와의 관계 정도를 탐구하고자 한다.

3 연구 방법

3.1 영어 말하기 데이터

본 연구는 영어 학습자 코퍼스 ICNALE(International Corpus of Asian Learners of English, Ishikawa 2019)에서 한국인 학습자의 영어 말하기 데이터를 추출하였다. ICNALE은 아시아 지역 여러 국가의 영어 학습자로부터 수집된 음성 발화 데이터와 문자 발화 데이터를 갖춘 방대한 학습자 코퍼스로서, 학습자 발화 분석에 널리 활용되고 있다(Dogar et al. 2024, Putra et al. 2021, Sung 2022). 본 연구는 음성 발화 데이터에 포함된 대화(dialogue) 코퍼스와 독백(monologue) 코퍼스 중 독백 코퍼스를 분석에 활용하였다.

ICNALE의 독백 코퍼스는 '식당에서의 흡연 금지'와 '대학생의 파트타임 근무'라는 두 개의 논증적 주제에 대한 말하기 데이터로 구성되어 있다. 모든 참가자들은 주제별로 2회씩, 총 4회에 걸쳐 말하기 과업을 수행하였다. 주제별 첫 번째 응답에는 20초의 준비 시간이, 두 번째 응답에는 10초의 준비 시간이 주어졌으며, 회당 60초 동안 음성 녹음이 진행되었다. 음성 녹취 파일은 전문 전사가에 의해 전사되었다. 모든 참여자는 IELTS, TOEFL, TOEIC, TEPS 등의 영어 능숙도 시험 점수 및 어휘 규모 시험(Nation and Beglar 2007) 결과에 따라 CEFR 기준으로 A2(waystage), B1_1(threshold: lower), B1_2(threshold: upper), B2+(vantage or higher) 중 하나의 능숙도 수준을 배정받았다.

한국인 학습자의 영어 독백 데이터는 100명의 대학생 학습자가 생산한 400개의 텍스트로 구성되어 있고, 영어 말하기 능숙도별 한국인 학습자 분포는 A2 6명, B1_1 15명, B1_2 43명, B2 36명이다. 본 연구는 능숙도에 따른 문법 다양성 척도를 분석하기 위하여, A2와 B1_1을 낮은 수준의 능숙도로, B2를 높은 수준의 능숙도로 설정하였고, 두 능숙도 그룹을 명확히 구분하기 위하여 그 사이에 해당하는 B1_2는 분석에서 제외하였다. 또한 말하기 연습 효과가 문법 다양성에 미치는 효과를 배제하기 위하여, 두 주제에 대한 첫 번째 녹음만을 분석 대상으로 설정하였다. 더불어 문법 다양성의 핵심 지표인 유형 빈도가 발화 길이에 영향을 받는다는 코퍼스 언어학의 기본 전제(McEnery and Hardie, 2011)에 입각하여 두 능숙도 집단의 평균 발화 길이가 유사하도록 학습자 발화 데이터를 선정하였다. 이를 위해 총 발화량이 100단어 초과 180단어 미만인 학습자 중에, A2 수준 학습자 5명은 전원 포함하였고, B1_1과 B2 수준은 각각 10명, 15명씩 무선표집하였다. 그 결과, 표 1과 같이 낮은 수준의 학습자 15명과 높은 수준의 학습자 15명이 두 주제에 대하여 최초 녹음에서 생산한 말하기 자료를 최종 분석 대상으로 선별되었다.

낮은 수준의 학습자 그룹은 A2 수준의 학습자 5명과 B1_1 수준의 학습자 10명으로 구성되어 있으며, 학습자 발화량의 범위는 105-179 단어, 평균은 135.9 단어였다. 이 수치는 높은 수준의 학습자 그룹과 유사하였는데, 높은 수준의 학습자는 발화량의 범위가 104-169 단어이고 평균이 139.1 단어였다. 독립 t-검정 결과, 두 집단은 발화량에서 유의미한 차이가 없었다($t(28) = -0.403$, $p = .690$, Cohen's $d = .147$).

표 1. 능숙도 집단별 학습자 데이터 구성

Low Proficiency Data			Higher Proficiency Data		
학습자ID	CEFR 수준	발화량(단어 수)	학습자ID	CEFR 수준	발화량(단어 수)
008	A2	109	007	B2	117
017	A2	179	026	B2	131
018	A2	159	028	B2	104
089	A2	164	033	B2	143
091	A2	112	042	B2	114
004	B1_1	157	044	B2	169
005	B1_1	131	047	B2	166
010	B1_1	146	048	B2	160
011	B1_1	134	055	B2	112
031	B1_1	131	056	B2	131
039	B1_1	119	061	B2	160
058	B1_1	105	063	B2	159
081	B1_1	131	064	B2	146
090	B1_1	145	067	B2	148
097	B1_1	116	100	B2	126
평균(범위)		135.9(105-179)	평균(범위)		139.1(104-169)

3.2 문법 다양성 지표 분석 도구

본 연구는 30명의 한국인 학습자 말하기 데이터에서 문법 다양성 지표 값을 여러 층위에서 도출하기 위하여 두 개의 자연어 처리 모듈을 사용하였는데, 하나는 Biber Tagger(Biber et al. 1999, Kyle et al. 2025)이고 다른 하나는 ASC annotator(Sung and Kyle 2024)이다. 첫 번째 모듈은 단어, 구, 절 다양성 분석에 활용되었고, 두 번째 모듈은 절 다양성 분석에 활용되었다. 두 모듈 모두 파이썬(Python)에 기반을 두고 있는데, 본 연구에서는 구글 코랩(Google Colab)의 환경을 이용하였다.

우선 본 연구에서 활용한 Biber Tagger는 파이썬 기반으로 재설계된 모듈로서, 기존의 Biber Tagger처럼 영어 텍스트를 자동으로 분석하고 그 안에 포함된 다양한 문법적 자질을 주석(annotation)으로 표시해 주는 소프트웨어이다. F1 점수는 0.857로서 전반적으로 우수한 수행을 보이는 언어 모듈이다. 이 모듈을 활용하기 위해, 우선 구글 코랩에서 자연어 처리 오픈소스 라이브러리인 'spaCy'와 영어 텍스트의 자연어 처리 작업을 수행하는 데 사용되는 Transformer 기반 언어 모델인 'en_core_web_trf'를 설치하였다. 이후 미국 오레곤 대학 LCR-ADS(Learner Corpus Research and Applied Data Science) 연구소의 github에 저장되어 있는 Biber tagger 모듈(2024년 9월 18일 기준)을 활용하여, 구글 드라이브에 저장되어 있는 한국인 영어 학습자 30인의 말하기 자료를 분석하였다. 예를 들어, "There are several reasons to support my idea"라는 학습자 발화를 위의 절차를 따라 처리하면 표 2와 같은 결과를 얻게 된다.

표 2. Biber Tagger 분석 예시

N	word	lemma	cx tag	main tag	xpos	deprel	headid
1	There	there			EX	expl	2
2	are	be		vbmain	VBP	ROOT	2
3	several	several	attr+ npremod	jj	JJ	amod	4
4	reasons	reason		nn	NNS	attr	2
5	to	to		to	TO	aux	6
6	support	support	tocls+ rel	vbmain	VB	relcl	4
7	my	my		dt	PRP$	poss	8
8	idea	idea		nn	NN	dobj	6

분석 결과의 특징을 살펴보면, 먼저 각 단어(word)를 행으로 구분하고 문장 내에서 순번(N)을 부여한다. 이후 단어의 원형(lemma), 문장 내 절 및 구의 정보(cx tag), 단어의 주요 분류(main tag), 최종 품사 태그(xpos), 의존 관계(deprel: dependency relation), 의존 관계에서 핵어의 위치(headid) 등의 정보가 자동으로 분석된다. 표 2에 나타나듯이, 각 층위별로 분석된 결과가 정확하고, 여러 수준의 언어 데이터가 종합적으로 활용되고 있어 안정적인 분석이 가능하다. 참고로, main tag 열과 xpos 열 사이에 단어의 세부 범주 정보(예: 수, 시제, 태, 격 등)에 관한 8개의 열도 있으나(cat1~8) 본 연구에서는 활용하지 않으므로 표 2에 제시하지 않았다. 본 연구는 위의 정보들을 측정 지표에 맞게 활용하여 단어, 구, 절 수준의 문법 다양성을 분석하였다.

본 연구의 두 번째 자연어 분석 모듈인 ASC Annotator는 텍스트 내 정형절(finite clause)과 비정형절(non-finite clause)이 어떠한 논항 구조 구문(ASC: argument structure construction)을 나타내는지를 분석하여, 절 수준의 문법 다양성 정보를 텍스트 단위로 산출한다(Sung and Kyle 2024). 예를 들어, “I kicked the ball”이라는 문장은 타동 구문(transitive construction)으로 분류되고, “I kicked him the ball”이라는 문장은 이중타동 구문(ditransitive construction)으로 분류된다. ASC Annotator는 Goldberg(1995)의 구문문법 체계에 입각하여, 표 3과 같이 총 9개의 구문을 분류 기준으로 삼고 있으며, L2 말하기 데이터에서 높은 정확도를 보인다(F1 = 0.928).

표 3. ASC Annotator의 논항 구조 구문 유형

논항 구조 구문	의미 구조	예시
Intransitive simple	agent-V	*I will be working this weekend*
Intransitive motion	theme-V-goal	*They went to the pool*
Intransitive resultative	patient-V-result	*The lake froze solid*
Transitive simple	agent-V-theme	*She cooked a pie*
Caused-motion	agent-V-theme-destination	*She put it on the table*
Transitive resultative	agent-V-result-result	*He kicked the door open*
Ditransitive	agent-V-recipient-theme	*I gave her a book*
Attributive	theme-V-attribute	*He is happy*
Passive	theme-aux-Vpassive	*The movie was underestimated*

3.3 문법 다양성 지표 분석 기준

Biber Tagger와 ASC Annotator에서 산출된 자연어 처리 데이터에 기반하여, 본 연구는 한국인 학습자의 영어 말하기 문법 다양성을 단어, 구, 절 수준에서 각각 두 종류의 유형 빈도(Type frequency)로 측정하였다. 단어 수준에서는 POS Tag 기준에 따른 품사 유형 빈도와 영어 교수-학습 기준에 따른 품사 유형 빈도를, 구 수준에서는 보편 의존관계 유형 빈도와 품사 이진 유형 빈도를, 절 수준에서는 종속절 유형 빈도와 논항 구조 구문 유형 빈도를 측정하였다.

3.3.1 단어 수준의 문법 다양성 측정

단어 수준의 문법 다양성은 수준을 달리하는 두 종류의 품사 유형 빈도로 측정하였다. 첫 번째 품사 유형 빈도는 Penn Treebank Project에 활용된 품사 구분을 기준으로 측정한 것으로, Biber Tagger 산출물의 xpos 열에 해당한다(표 2 참고). Penn Treebank의 품사 태그는 표 4에 제시된 것과 같이 총 36개인데, 이 가운데 FW(Foreign word, 예: per), POS(Possessive ending, 예: 's), SYM(Symbol, 예: '...'), CD(Cardinal number, 예: 1,000) 등 4개의 태그는 영어 단어 발화와 관련이 적으므로 유형 빈도 분석에서 제외하였다. 품사 유형 빈도는 학습자 단위로 산출되었는데, 두 주제에 대한 전체 발화에서 1회 이상 사용된 품사의 유형을 빈도화하였다.

표 4. Biber Tagger에 사용된 Penn Treebank 품사 태그

태그	설명	태그	설명	태그	설명
CC	Coordinating conjunction	NNS	Noun, plural	TO	to
CD	Cardinal number	NNP	Proper noun, singular	UH	Interjection
DT	Determiner	NNPS	Proper noun, plural	VB	Verb, base form
EX	Existential there	PDT	Predeterminer	VBD	Verb, past tense
FW	Foreign word	POS	Possessive ending	VBG	Verb, -ing form
IN	Preposition/subordinator	PRP	Personal pronoun	VBN	Verb, past participle
JJ	Adjective	PRP$	Possessive pronoun	VBP	Verb, present
JJR	Adjective, comparative	RB	Adverb	VBZ	Verb, 3rd singular
JJS	Adjective, superlative	RBR	Adverb, comparative	WDT	Wh-determiner
LS	List item marker	RBS	Adverb, superlative	WP	wh-pronoun
MD	Modal	RP	Particle	WP$	Possessive wh-pronoun
NN	Noun, singular or mass	SYM	Symbol	WRB	Wh-adverb

두 번째 품사 유형 빈도는 영어 교수-학습에서 활용되는 품사 구분을 기준으로 측정하였다. 이를 위해, 2022 개정 영어과 교육과정(교육부 2022)과 영문법 전문 서적(예: 문용 2017, Cowan 2008)을 참고하여 표 5와 같이 Biber Tagger의 POS Tag 값을 영어 교수-학습에서 통용되는 13개의 품사 범주로 재구조화하였다. 유형 빈도 산출은 마찬가지로 학습자 단위로 산출되었으며, 1회 이상 사용된 품사의 유형 빈도를 측정하였다.

표 5. 영어 교수-학습에서 활용되는 품사 유형

교수-학습 기준 품사	Biber Tagger POS Tag 값
명사	NN, NNS, NNP, NNPS
대명사	PRP, PRP$
동사	VB, VBD, VBG, VBN, VBP, VBZ
부사	EX, RB, RBR, RBS, RP
형용사	JJ, JJR, JJS
전치사	IN 중 대부분
접속사	IN 중 일부, CC
한정사	DT, PDT
관계사	WDT
조동사	MD
TO	TO
감탄사	UH
의문사	WRB, WP

3.3.2 구 수준의 문법 다양성 측정

구 수준의 문법 다양성 또한 두 종류의 유형 빈도로 측정하였는데, 하나는 품사 이진 유형 빈도이고, 다른 하나는 보편 의존관계(Universal Dependencies) 유형 빈도이다.

우선, 품사 이진은 연속하는 2개 단어의 품사 배열 패턴을 의미하는데, 영어 자동 평가에서 널리 사용되는 자질로서 학습자의 능숙도에 따라 자주 사용되는 품사 배열 패턴이 다른 것으로 알려져 있다(Zechner and Evanini 2019). 예를 들어, “Maria has left a note.” 문장을 Penn Treebank 기준으로 품사 태깅하면 “Maria_NNP has_VBZ left_VBN a_DT note_NN ._.”의 결과가 나온다. 여기에서 인접한 품사 태그를 2개씩 묶으면 “NNP-VBZ, VBZ-VBN, VBN-DT, DT-NN, NN-.”이 되는데, 이 중 구두점을 포함하고 있는 마지막 사례를 제외하면 총 네 유형의 이진(Bigram)이 남는다. 본 연구는 이와 같은 방법으로 학습자별 1회 이상 사용된 품사 이진 유형을 빈도화 하였는데, 이를 위해 Biber Tagger 산출물에서 xpos 열에 제시된 품사 태그 정보를 활용하였다.

반면, 보편 의존관계는 100여 개의 자연어에 보편적으로 적용되는 형태-통사 주석(morphosyntactic annotation) 체계로서, 문장 구조에서 핵어(head)와 의존어(dependent)의 관계를 정의한다(De Marneffe et al. 2021). 예를 들어, 문장 “Maria has left a note.”의 보편 의존관계는 그림 1과 같고, 의존관계는 핵어에서 의존어로 향하는 화살표와 의존관계 자질(둥근 사각형 정보)로 표시된다.

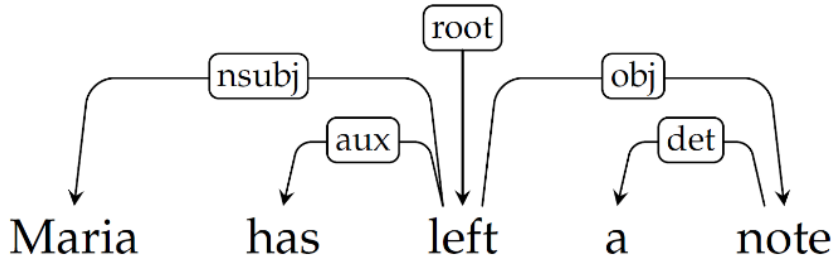

그림 1. 보편 의존관계 사례 도식(De Marneffe et al. 2021, p. 273)

분석 결과를 살펴보면, 동사 left는 문장 전체의 핵어인 root로서 의존어 Maria, has, note와 각각 주어(nsubj), 조동사(aux), 목적어(obj)의 관계를 가진다. 또한 명사구 a note에서 핵어인 note는 의존어 a와 한정어(determiner)의 관계를 가진다. 이처럼 문장의 모든 단어는 의존관계 자질을 가지게 되는데, 둥근 네모 상자에 제시된 정보를 참고하면 이 문장의 보편 의존관계 유형 빈도는 5이다(i.e., nsubj, aux, root, det, obj). 본 연구는 Biber Tagger 산출물의 deprel 열에 제시된 보편 의존관계 주석 정보를 활용하여, 학습자별 1회 이상 사용한 의존관계 자질을 유형 빈도화 하였다.

3.3.3 절 수준의 문법 다양성 측정

절 수준의 문법 다양성은 종속절 유형 빈도와 논항 구조 구문 유형 빈도로 측정하였다. 종속절 유형 빈도 측정을 위해, Biber Tagger 산출물에서 cx tag 열의 정보를 활용하였다. 해당 열은 종속절의 형태-의미적 자질 정보를 제공하는데, 본 연구는 형태적 자질만을 고려하여, 표 6과 같이 4개의 정형절과 3개의 비정형절을 포함하는 총 7개 범주로 유형화하고, 학습자별 1회 이상 사용된 종속절 범주의 유형 빈도를 측정하였다.

표 6. 형태적 자질을 반영한 종속절 범주

범주	학습자 발화 예시
정형_접속부사	... for university students *because I can experience social jobs.*
정형_관계사	it is good experience for the people *who got part-time job.*
정형_의문사	keep learning about *what they really like.*
정형_that	I think *that cigarette of smoking is very harmful to the body.*
비정형_ing	And without any job it's too hard in *living in Seoul.*
비정형_to	we can't give the opportunity *to have the part-time job.*
비정형_ed	*Given the long-term visions*, the semester is not enough.

또 다른 절 수준 문법 다양성 척도인 논항 구조 구문 유형 빈도는, ASC Annotator의 산출물을 기준으로 각 학습자가 9개 유형의 ASC 중 몇 개의 유형을 1회 이상 사용했는지 측정하였다(표 3 참조).

3.4 통계 분석

본 연구는 Jamovi 2.6.26을 활용하여, 한국인 학습자 30인의 말하기 데이터에 대한 상관분석과 회귀분석을 실시하였다. 상관 분석은 6개의 문법 다양성 지표 간에 실시하였는데, 모든 지표는 유형 빈도에 해당하므로 정수 형식의 연속 변인으로 설정하였다. 분석에는 Pearson Correlation Coefficients가 사용되었으며 알파 값은 .05로 설정하였다.

이후 이항형 로지스틱 회귀분석(Binomial Logistic Regression)을 실시하여 영어 학습자의 능숙도에서 문법 다양성 지표 값의 설명력을 확인하였다. “로지스틱 회귀분석”은 종속 변인이 연속형 변인(예: 무게)이 아니고 범주형 또는 순서형 변인(예: 상-중-하)일 때 사용되며, 순서형 범주가 두 개일 때는 “이항형” 분석이 실시된다. 본 연구의 종속 변인인 능숙도는 1(낮은 수준)과 2(높은 수준) 값의 순서형 변인으로 설정되었다. 모델 설계는 종속 변인과의 상관관계가 높은 독립 변인부터 추가하는 방식의 전진 선택법(forward-stepwise)을 따랐으며, 모델의 적합도와 설명력은 각각 AIC(Akaike Information Criterion)와 McFadden R^2로 검증하였다. 독립 변인 간의 다중공선성(multicollinearity) 문제를 예방하기 위하여 상관 계수를 활용하였는데, 전진 선택법 분석에서 기존 변인 간의 상관 계수가 0.7 미만인 변인만을 추가하였다.

4. 연구 결과 및 논의

4.1 문법 다양성 지표 간 상관 분석

한국인 학습자의 영어 말하기 데이터에 대해 단어, 구, 절 단위에서 각각 2개의 문법 다양성 지표의 유형 빈도를 분석한 결과, 표 7과 같은 분포를 보였다. 우선 단어 수준의 지표를 비교했을 때, Penn Treebank 품사 태그 유형과 영어 교수-학습 맥락의 품사 유형이 범위에서 큰 차이를 보였다. 예를 들어, 32개의 Penn Treebank 품사 태그를 모두 사용한 학습자는 없었으나, 13개의 교수-학습 맥락의 품사 유형을 모두 사용한 학습자는 있었다. 구 수준의 지표를 살펴보면, 품사 이진 유형 빈도가 매우 큰 범위를 보였고, 보편 의존 유형도 최소 20개 이상으로 다양하게 나타났다. 절 수준의 지표를 보면, 종속절 7개 유형과 논항 구조 구문 9개 유형을 모두 사용한 학습자는 없었고, 평균 3~4개의 종속절 및 논항 구조 구문 유형을 활용한 것으로 나타났다.

표 7. 문법 다양성 지표의 유형 빈도

수준	지표	평균	표준편차	최소-최대값
단어	NLP 태그 품사	18.0	2.7	13-23
	교육용 태그 품사	11.4	1.0	10-13
구	품사 이진	53.7	12.4	35-79
	보편 의존	26.2	3.0	20-32
절	종속절	3.8	1.3	1-6
	논항 구조	4.2	1.0	3-6

이상의 6개의 문법 다양성 지표 유형 빈도 값에 대하여 상관관계 분석(Pearson r)을 시행한 결과, 표 8과 같았다. 분석 결과를 살펴보면, 우선 대부분의 지표 쌍이 유의미한 상관관계를 보인다. 가령, 단어 수준 다양성의 두 지표인 Penn Treebank NLP 기준 품사 유형 빈도와 영어 교수-학습 기준 품사 유형 빈도는 높은 상관관계(r = .738, p < .001)를 보이고 있고, 구 수준 다양성 지표인 품사 이진과 보편 의존 빈도 역시 비슷한 수준의 높은 상관관계(r = .699, p < .001)를 보인다. 이는 수준별 두 가지 지표가 유사한 능력을 측정할 수 있음을 의미하며, 따라서 말하기 자동 평가에서 두 가지 지표를 함께 포함할 경우, 유사한 능력을 중복 배점하는 문제가 발생할 수 있다. 반면, 절 수준의 문법 다양성에 해당하는 종속절 유형 빈도와 논항 구조 유형 빈도는 유의미한 상관관계를 보이지 않았다(r = .263, p = .161). 이는 한국인 학습자의 영어 말하기 데이터에서 종속절의 유형과 논항 구조 구문의 유형이 독립적인 양상을 보임을 의미한다. 즉, 다양한 종속절 유형을 사용하는 학습자가 반드시 다양한 논항 구조 구문을 사용할 것으로 예측할 수 없다. 따라서 한국인 영어 학습자가 두 영역 모두에서 다양성을 계발하고자 한다면, 각 영역에 특화된 노력이 필요할 것이다.

표 8. 문법 다양성 지표 간 상관관계

지표	NLP 태그 품사	교육용 태그 품사	품사 이진	보편 의존	종속절	논항 구조
NLP 태그 품사	—					
교육용 태그 품사	.738***	—				
품사 이진	.875***	.736***	—			
보편 의존	.713***	.555**	.699***	—		
종속절	.717***	.625***	.595***	.349	—	
논항 구조	.563**	.458*	.605***	.575***	.263	—

비고: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

표 8의 상관 분석 결과의 또 다른 특징은, 단어 수준의 품사 관련 지표들이 구나 절 수준의 지표들과 유의미한 상관관계를 보인다는 점이다. 특히 Penn Treebank 기반 자연어처리로 측정한 품사 유형 빈도와 품사 배열에서 추출한 품사 이진 유형 빈도는 가장 높은 수치의 상관관계(r = .875, p < .001)를 보였다. 이러한 결과는, 한국인 학습자 영어 말하기에서 품사 유형이 다양해질수록 구 구조와 절 구조의 유형이 함께 다양해짐을 의미한다.

4.2 문법 다양성 지표와 능숙도 간 회귀분석

한국인 학습자의 영어 능숙도 집단별로 문법 다양성 지표 분포를 살펴본 결과, 표 9와 같았다. 대부분의 지표는 상위 수준 학습자 그룹에서 더 높은 평균값을 보였다. 그러나 논항 구조 구문 유형 빈도는 반대의 양상을 보이며, 능숙도와 부적 상관관계를 가지는 것으로 밝혀졌다. 즉, 본 연구에서 분석한 말하기 데이터에서는 논항 구조의 다양성이 증가하면 낮은 능숙도에 속할 경향이 증가하였다. 이는 능숙도가 발달할수록 문법 다양성이 증가한다는 본 연구의 전제에 반대되는 내용이므로, 해당 지표는 회귀분석에서 제외되었다.

표 9. 한국인 학습자의 영어 능숙도별 문법 다양성 지표의 유형 빈도

수준	지표	낮은 능숙도 집단		높은 능숙도 집단		능숙도와 상관관계
		Min-Max	Mean	Min-Max	Mean	
단어	NLP 태그 품사	13-21	17.20	15-23	18.87	0.314
	교육용 태그 품사	10-13	11.07	11-13	11.67	0.316
구	품사 이진	36-79	51.87	35-79	55.60	0.153
	보편 의존	20-30	25.80	22-32	26.67	0.147
절	종속절	1-5	3.13	2-6	4.47	0.534
	논항 구조	3-6	4.33	3-6	4.13	-0.105

능숙도와의 상관관계 수치를 기준으로 전진 선택형 이항형 로지스틱 회귀분석을 실시한 결과, 표 10에 제시된 바와 같이, 종속절 지표로만 설계된 Model 1은 통계적으로 유의하였고($\chi^2(1) = 9.74$, $p < .01$) 약 23.4%(McFadden R^2)의 추정 설명력을 보였다. 그러나 후속 모델은 어떠한 변인을 삽입해도, 설명력은 일부 증가했으나 AIC 값이 커지면서 모델 적합도가 감소하는 경향을 보였다. 이 가운데 품사 이진 지표를 추가한 경우 모델 적합도의 감소 폭이 가장 작고 설명력의 증가 폭이 가장 컸기 때문에, 본 연구는 이를 Model 2로 제시하였다. Model 2 또한 통계적으로 유의하였고($\chi^2(2) = 11.60$, $p < .01$), 약 27.9%의 추정 설명력을 보였다.

표 10. 전진 선택형 이항형 로지스틱 회귀분석 결과

Model	Deviance	AIC	BIC	R^2McF	χ^2	df	*p*
1	31.8	35.8	38.7	0.234	9.74	1	.002
2	30.0	36.0	40.2	0.279	11.60	2	.003

모델별 변인의 설명력은 표 11과 같다. 우선 Model 1에서는 종속절 유형 빈도가 통계적으로 유의하였고($p = .01$) 해당 지표의 계수 추정값은 1.13(odds ratio = 3.09)으로 나타났다. 이는 종속절 유형 빈도가 1단위 증가할 때, 학습자가 높은 능숙도를 보일 가능성이 약 3.1배 증가함을 의미한다. Model 2에서도 종속절 유형 빈도는 통계적으로 유의하였고($p = .012$), 지표의 계수 추정값은 1.6(Odds ration = 4.99)으로, 종속절의 유형 빈도가 1 증가할 때 학습자가 높은 능숙도를 보일 가능성이 약 5배 증가하는 것으로 나타났다. 반면 품사 이진 유형 빈도는 통계적으로 유의하지 않았다($p = .202$).

표 11. 모델별 변인 회귀계수

Model	Predictor	Estimate	95% Conf. Interval		SE	Z	*p*	Odds ratio
			Lower	Upper				
1	절편	-4.32	-7.75	-0.88	1.75	-2.46	.014	0.01
	종속절	1.13	0.27	1.99	0.44	2.57	.010	3.09
2	절편	-2.66	-6.84	1.53	2.14	-1.24	.214	0.07
	종속절	1.61	0.35	2.86	0.64	2.51	.012	4.99
	품사 이진	-0.07	-0.17	0.04	0.05	-1.27	.202	0.94

표 10과 표 11에 제시된 결과를 종합하면, 한국인 영어 말하기 데이터에서 종속절 유형의 다양성이 모든 통계 모델에서 영어 능숙도를 유의미하게 설명하였다. Model 1은 종속절 유형의 다양성이라는 단일 변인을 중심으로 비교적 간결한 구조를 갖추고 있으며, AIC 기준(35.8)에서 가장 낮은 값을 보여 통계적 경제성이 뛰어난 모델로 볼 수 있다. Model 2는 품사 이진 유형 빈도가 추가되었는데, 종속절 유형 빈도는 여전히 유의한 변인으로 나타났다. 흥미롭게도, 품사 이진 유형 빈도가 추가되면서 종속절 유형 빈도의 추정 계수는 1.13에서 1.60으로 상승하였다. 이는 품사 이진 유형 빈도가 통제된 상태에서 종속절 다양성의 예측력이 더 명확하게 드러났음을 시사한다. 한편, 품사 이진 유형 빈도의 추정 계수는 유의하지 않았지만, 전체 모델의 설명력은 McFadden R^2 기준으로 .234에서 .279로 증가하였다. AIC 값은 36.0으로 소폭 상승하였으나(Δ = 0.2), 이는 통계적으로 의미 있는 수준의 적합도 하락으로 보기는 어렵다. 이러한 점에서 Model 2는 통계적 간결성은 다소 낮지만, 설명력 제고와 해석의 풍부함 측면에서 분석적 가치를 지닌다고 볼 수 있다.

마지막으로 영어 능숙도에 대한 종속절 유형 빈도의 설명력을 탐구하고자, 종속절 유형별 사용 학습자 수를 능숙도 집단 간 비교하였다(표 12 참조). 우선 두 집단 모두 대부분의 학습자가 “정형_접속부사” 종속절(예: because I ...)을 사용했고, 매우 적은 수의 학습자가 “비정형_ed” 종속절(예: given the ...)을 사용했다. 두 집단 간 주된 차이는 “정형_관계사(예: people who ...)”, “비정형_ing(예: living in ...)”, “비정형_to(예: to have ...)” 등 세 유형의 종속절에서 발견되었다.

표 12. 종속절 유형별 사용 학습자 수 그룹 간 비교

종속절	정형_ 접속부사	정형_ 관계사	정형_ 의문사	정형_ that	비정형_ ing	비정형_ to	비정형_ ed
상위 그룹	15	9	4	13	11	13	2
하위 그룹	13	4	2	14	5	8	1
차이	2	5	2	-1	6	5	1

5. 결론

본 연구는 한국인 영어 학습자의 말하기 평가에 활용될 수 있는 문법 다양성 지표를 단어, 구, 절 수준에서 광범위하게 탐색하였다. 말하기 자동 평가의 기술 동향과 제2언어 습득 및 학습 연구 등에 입각하여 각 수준별로 2개의 지표를 설정하고, 파이썬 기반 자연어 처리 기술을 활용하여 체계화된 지표 값을 산출하였다. 두 개의 연구 질문에 근거하여 지표 간 상관관계와 학습자 영어 능숙도와의 관련성을 분석한 주요 결과는 다음과 같다.

우선 문법 다양성 지표 간 상관관계 분석에서 대부분의 지표 쌍에서 유의미한 상관관계를 발견했다. 또한 단어 수준의 품사 정보에 근거한 유형 빈도 지표들이 구와 절 수준의 모든 지표들과 유의미한 상관관계를 보였다. 이는 단어 수준의 문법 다양성 지표가 구나 절 수준의 문법 다양성과 연계되어 있음을 의미하며, 따라서 한국인 영어 말하기 자동 평가 시스템에 품사 기반 지표를 포함할 경우, 이와 같은 상관관계를 고려하여 중복 평가가 되지 않도록 평가 모델과

수식을 설계할 필요가 있음을 시사한다. 나아가 품사 유형 빈도가 다른 문법 다양성 지표들과 높은 상관관계를 보이는 구체적인 양상을 분석하여, 그 결과를 한국인의 영어 학습 및 평가를 고도화하는 데 활용할 수 있을 것이다.

다음으로 문법 다양성 지표와 학습자 영어 능숙도와의 관련성을 이항형 로지스틱 회귀분석으로 분석하여 두 개의 통계 모델을 도출하였다. Model 1은 적합성과 경제성을 우선시하는 모델로서, 종속절 유형 빈도를 단일 변인으로 설정하였다. 이 모델은 변수 간의 상호작용을 고려하지 않아 해석이 용이하고 직관적이다. 그러나 McFadden R^2 값에서 나타나듯이 모델이 제공할 수 있는 설명이 제한적이기에, 다른 변수들이 추가된다면 더 나은 설명력을 얻을 수 있을 가능성이 존재한다. 반면, Model 2는 품사 이진 유형 빈도를 추가함으로써 모델의 설명력은 향상되었으나, 모델의 적합도가 소폭 감소하였고 복수 변인으로 인해 해석이 다소 복잡해졌다. 따라서 두 모델은 각기 다른 교육적 요구에 맞춰 선택될 필요가 있다. Model 1은 간결하고 빠른 분석과 직관적인 피드백을 제공하여 교육적 활용이 용이하다. 반면, Model 2는 정확한 분석을 통해 학습자의 능숙도를 심층적으로 평가하고, 그 결과를 중요한 결정에 활용하고자 할 때 적합하다.

로지스틱 회귀분석 모델의 변인별 추정 계수를 분석한 결과, 6개의 문법 다양성 지표 중 절 수준에서 측정된 종속절 유형 빈도만이 영어 능숙도를 유의미하게 설명하는 것으로 나타났다. 이는 종속절 활용이 영어 말하기 평가에서 중요한 지표로 활용될 수 있다는 해외 연구자들의 주장을 뒷받침한다(Biber et al. 2011, Kim and Lu 2024, Vercellotti, 2019). 이러한 결과는 논증적 말하기(argumentative speaking)라는 말하기 과업 유형의 특징과도 관련이 있어 보인다. 의견을 개진하고 인과관계를 설명하고 상황을 제시하는 등 논증적 말하기의 주요 기능에서 다양한 유형의 종속절이 활용되는데, 이러한 종속절의 사용 정도가 학습자 능숙도에 따라 달라질 수 있다. 물론 다른 과업 유형을 사용한 경우에도 평가 지표의 설명력이 동일한 양상을 보이는지에 대해서는 추가 연구가 필요하다. 만일 과업 유형에 따라 평가 지표의 설명력이 달라진다면, 과업 유형별로 평가 지표를 다르게 활용하는 평가 모델을 개발할 필요가 있을 것이다.

반면 상당 수의 영어 자동 평가 시스템에서 활용하고 있는 품사 이진 유형 빈도는 본 연구에서 영어 능숙도와 직접적인 유의미한 관계를 보이지 않았다. Model 2에서 품사 이진 유형 빈도를 추가하였을 때 모델의 전체적인 설명력과 종속절 유형 빈도의 추정 계수는 상승하였지만, 해당 변인 자체는 유의미한 설명력을 가지지 못했다. 이는, 논증적 말하기 데이터의 문법 다양성으로 영어 능숙도를 예측할 때 품사 이진 유형 빈도가 예측 변인보다는 통제 변인으로 활용될 수 있음을 시사한다. 나아가 최종 설명 모델에 포함되지 못하였거나 제외된 지표들의 경우, 다른 학습자 집단과 다른 말하기 과업에서도 동일한 결과를 보이는지 후속 연구로 확인될 필요가 있다.

한국인 학습자의 영어 말하기 자동 평가 개발을 위한 문법 다양성 지표를 단어, 구, 절의 다층위에서 탐색한 본 연구는, 후속 연구에서 보완되어야 할 여러 제한점을 지니고 있다. 첫째, 참여자 구성이 A2~B2 수준의 30인의 한국인 대학생들로 제한되어 있는데, 후속 연구에서는 분석 규모를 확대하여 연령대와 수준이 다양한 학습자를 탐구할 필요가 있다. 예를 들어, C1 수준의 학습자나 영어 원어민 데이터를 함께 분석하여, 높은 수준의 학습자와 원어민을 판별하는 데 동일한 지표가 적용 가능한지 확인할 수 있다. 둘째, 본 연구에서는 논증적 말하기 데이터만을 분석하였는데, 대화하기와 묘사하기 등 다른 주요 유형의 말하기 데이터를 대상으로 한 연구가 필요하다. 셋째, 문법 다양성과 함께 유창성, 발음, 내용, 어휘 등 말하기 능숙도에 영향을 미치는

다양한 변인들을 종합적으로 고려한 연구가 필요하다. 마지막으로 본 연구의 능숙도 데이터가 참여자의 영어 말하기 능숙도가 아닌 일반적인 영어 능숙도를 나타내는 점 역시 후속 연구에서 보완이 될 필요가 있다. 이러한 노력을 지속적으로 기울인다면 한국인 영어 말하기 평가 시스템의 채점 신뢰도와 교육으로의 기여도를 확보할 수 있을 것이다.

참고 문헌

교육부(Ministry of Education). 2022. 『2022 개정 영어과 교육과정』(2022 revised national curriculum for English). 교육부(Ministry of Education).

남보라·조규희·박지현·성민창·황필아·강정진·이동환·심창용·최희경·박선호·김혜련(Nam, B., K. Jo, J. Park, M. Sung, P. Hwang, J. Kang, D. Lee, C. Sim, H. Choi, S. Park and H. Kim). 2024. 『인공지능 융합 영어교육의 이해』(*Understanding of AI convergence English education*). 경문사(Kyungmoonsa).

문용(Moon, Y.). 2017. 『고급 영문법 해설』(*Advanced English grammar explanation*). 박영사(Parkyoungsa).

신동광·박용효·박태준·임수연(Shin, D, Y. Park, T. Park and S. Yim). 2015. 영어 말하기 자동채점의 현재와 미래(The present and future of an automated scoring program for speaking skills of English). ≪멀티미디어 언어교육≫(*Multimedia-Assisted Language Learning*) 18(1), 93-114.

이진화·최윤덕·성민창·김혜영(Lee, J.-H., Y. Choi, M. Sung and H. Kim). 2023. 자동채점 기반 영어 말하기 시험 현황 분석(Analysis of English automated speaking scoring tests). ≪영어교육≫(*English Teaching*) 78(2), 223-244.

ACTFL. 2012. *ACTFL proficiency guidelines 2012.* ACTFL.

ACTFL. 2024. *ACTFL proficiency guidelines 2024.* ACTFL.

Bhat, S. and S. Y. Yoon. 2015. Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication* 67, 42-57.

Bhat, S., H. Xue and S. Y. Yoon. 2014. Shallow analysis based assessment of syntactic complexity for automated speech scoring. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1305-1315.

Biber, D., B. Gray and K. Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45, 5-35.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan. 1999. *Longman grammar of spoken and written English.* Longman.

Chen, L., K. Zechner, S. Yoon, K. Evanini, X. Wang, A., Loukina, … and B. Gyawali. 2018. *Automated scoring of nonnative speech using the SpeechRater v. 5.0 engine.* (ETS Research Report No. RR-18-10). Educational Testing Service. Available online at

https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12198

Choi, J. and M. Sung. 2020. Utterance-based measurement of L2 fluency in speaking interactions: A constructionist approach. *English Teaching* 75(S1), 105-126.

Council of Europe. 2020. *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume.* Council of Europe Publishing.

Cowan, R. 2008. *The teacher's grammar of English with answers: A course book and reference guide*. Cambridge University Press.

De Marneffe, M. C., C. Manning, J. Nivre and D. Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2), 255-308.

Dogar, M. F., T. Saleem, M. Aslam, and S. Khan. 2024. Exploring global linguistic nuances: Analyzing region-specific inflectional morpheme frequency in ICNALE. *Asian-Pacific Journal of Second and Foreign Language Education* 9(1), 65.

Dulay, H. C. and M. K. Burt. 1974. Natural sequences in child second language acquisition. *Language Learning* 24(1), 37-53.

Ellis, N. C. and F. Ferreira-Junior. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal* 93(3), 370-385.

Goldberg, A. E. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Ishikawa, S. 2019. The ICNALE spoken dialogue: A new dataset for the study of Asian learners' performance in L2 English interviews. *English Teaching* 74(4), 153-177.

Khristoforov, S., V. Bochkarev and A. Shevlyakova. 2020. Recognition of parts of speech using the vector of bigram frequencies. In W. van der Aalst et al., eds., *Analysis of Images, Social networks and Texts: AIST 2019,* 137-147. Springer.

Kim, M. and X. Lu. 2024. L2 English speaking syntactic complexity: Data preprocessing issues, reliability of automated analysis, and the effects of proficiency, L1 background, and topic. *The Modern Language Journal* 108(1), 270-296.

Krashen, S. and T. Terrell. 1983. *The natural approach: Language acquisition in the classroom.* Pergamon Press.

Kyle, K. 2016. *Measuring Syntactic Development in L2 writing: Fine-grained Indices of Syntactic Complexity and Usage-based Indices of Syntactic Sophistication.* Doctoral dissertation, Georgia State University.

Kyle, K. and S. Crossley. 2017. Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing* 34(4), 513-535.

Kyle, K., H. Sung, H., D. Biber, R. Reppen, R. and J. Egbert. 2025. *The development and evaluation of an open-source lexicogrammatical complexity analysis tool: The Lexicogrammatical Tagger.* Paper presented at the Annual Conference of American Association of Applied Linguistics.

Lee, J.-H. and H. M. Kim. 2011. The L2 developmental sequence of English constructions and

underlying factors. *Korean Journal of English Language and Linguistics* 11(3), 577-600.

McEnery, T. and A. Hardie. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Nation, I. S. P. and D. Beglar. 2007. A vocabulary size test. *The Language Teacher* 31(7), 9-13.

Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24(4), 492-518.

Park, J. and M. Sung. 2024. Expansion of verb-argument construction repertoires in L2 English writing. *International Review of Applied Linguistics in Language Teaching* 62(2), 903-925.

Putra, J. W. G., S. Teufel and T. Tokunaga. 2021. *Multi-task and multi-corpora training strategies to enhance argumentative sentence linking performance.* arXiv preprint arXiv:2109.13067.

Skehan, P. 1996. A framework for the implementation of task-based instruction. *Applied linguistics* 17, 38-62.

Sung, H. and K. Kyle. 2024. Leveraging pre-trained language models for linguistic analysis: A case of argument structure constructions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7302-7314.

Sung, M. 2022. Lexical verb forms in L1 and L2 spoken English: A corpus-based analysis. *The Journal of Education* 42(S), 53-66.

Sung, M. and H. Kim. 2022. Effects of verb-construction association on second language constructional generalizations in production and comprehension. *Second Language Research* 38(2), 233-257.

Vercellotti, M. L. 2019. Finding variation: Assessing the development of syntactic complexity in ESL Speech. *International Journal of Applied Linguistics* 29, 233-247.

Xu, J., E. Jones, V. Laxton and E. Galaczi. 2021. Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. *Assessment in Education: Principles, Policy & Practice* 28(4), 411-436.

Yoon, S. Y. and S. Bhat. 2012. Assessment of ESL learners' syntactic competence based on similarity measures. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 600-608.

Zechner, K. and K. Evanini. 2019. *Automated speaking assessment: Using language technologies to score spontaneous speech.* Routledge.

예시 언어(Examples in): English

적용 가능 언어(Applicable Languages): English

적용 가능 수준(Applicable Level): Secondary & Tertiary

