The Journal of the Convergence on Culture Technology (JCCT)
Vol. 8, No. 5, pp.691-696, September 30, 2022. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2022.8.5.691
JCCT 2022-9-86

판별분석을 통해 살펴본 영어 능력 수준을 구별하는 어휘의 정교화 특성

Lexical Sophistication Features to Distinguish the English Proficiency Level Using a Discriminant Function Analysis

이영주*

Young-Ju Lee*

요 약 본 연구는 영어 능력 수준을 구별할 수 있는 어휘적 정교화 특징이 무엇인지를 자동화된 어휘 분석 프로그램인 TAALES를 활용하여 탐색하였다. 300명의 한국 대학생이 쓴 총 600개의 에세이가 ICNALE 코퍼스에서 추출되었고 SPSS 프로그램의 판별 분석이 수행되었다. 판별 분석 결과 한국 대학생을 상. 중. 하의 세 개의 영어 능력 수준으로 유의미하게 구분하는 어휘 특성은 SUBTLEXUS 코퍼스의 내용어 빈도, 내용어의 어휘 습득 연령, 기능어의 어휘 결정 반응 평균 시간, 상위어 동사로 나타났다. 영어 능력 수준이 높은 상 수준 학생은 SUBTLEXUS 코퍼스에 빈번하게 나오는 어휘는 많이 사용하지 않았고, 어휘 습득 연령이 높고 어휘 결정 과업에서 평균 반응시간이 길게 나타난 정교화된 어휘와 구체적인 동사를 많이 사용한 특징이 있다.

주요어 : 판별분석, 어휘의 정교화, TAALES 프로그램, 영어 능력 수준

Abstract This study explored the lexical sophistication features to distinguish the group membership of English proficiency, using the automatic analysis program of lexical sophistication. A total of 600 essays written by 300 Korean college students were extracted from the ICNALE (International Corpus Network of Asian Learners of English) corpus and a discriminant function analysis was performed using SPSS program. Results showed that the lexical features to distinguish three groups of English proficiency are SUBTLEXUS frequency content words, age of acquisition content words, lexical decision mean reaction time function words, and hypernymy verbs. High-level Korean students used frequent content words from SUBTLEXUS corpus to a lesser degree and produced more sophisticated words that can be learned at a later age and take longer reaction time in lexical decision task, and more concrete verbs.

Key words : A Discriminant Function Analysis, Lexical Sophistication, The Tool for the Automatic Analysis of Lexical Sophistication (TAALES), English Proficiency Level

*정회원, 한밭대학교 영어영문학과 교수 (주저자)
접수일: 2022년 8월 19일, 수정완료일: 2022년 8월 28일
게재확정일: 2022년 9월 9일

Received: August 19, 2022 / Revised: August 28, 2022
Accepted: September 9, 2022
*Corresponding Author: yjulee@hanbat.ac.kr
Professor, Dept. of English Language and Literature, Hanbat National University, Korea

Ⅰ. 서 론

본 연구에서는 자동화된 어휘 분석 프로그램인 TAALES (the Tool for the Automatic Analysis of Lexical Sophistication)를 활용하여 한국 대학생을 상, 중, 하의 세 개의 영어 수준으로 분류할 수 있는 어휘의 정교화 특성에는 무엇이 있는가를 살펴보고자 한다. TAALES 프로그램에서 어휘의 정교화는 크게 5개의 요소로 구성되고 어휘의 빈도(frequency), 어휘의 범위(range), 한 단어(unigram), 두 단어(bigram), 세 단어(trigram) 조합의 n-gram, 학문어휘 리스트(academic word list), 심리언어학적 특성이 포함된다 [1]. Coh-Metrix 프로그램에서는 어휘의 범위, n-gram, 학문어휘 리스트와 같은 지표는 분석할 수 없으나, TAALES 프로그램에서는 분석이 가능하다는 장점이 있다.

Ⅱ. 선행연구 검토 및 비교

어휘의 정교화는 Coh-Metrix 프로그램을 활용하여 국내외에서 에세이를 분석한 많은 선행연구[2-6]가 수행되었다. Crossley와 McNamara는 홍콩의 고등학교 졸업생이 작성한 에세이의 어휘의 정교화를 Coh-Metrix 프로그램을 활용하여 분석하였다 [3]. 분석결과를 요약하면, 어휘의 다양성, 단어 친숙도(word familiarity), 내용어 빈도(content word frequency), 어휘의 의미성(word meaningfulness), 상의 반복(aspect repetition)의 5개 변인은 훈련용 데이터(training set)의 경우 에세이 점수 변인의 약 29%를 차지하는 것으로 나타났다. 5개의 변인 중에서 어휘의 다양성은 에세이 점수 변인의 18%를 차지하며 에세이 점수를 예측하는 가장 강력한 예측변인으로 나타났다. 즉, 상위수준의 학생일수록 다양하고 친숙도와 빈도가 낮으며 다른 단어와의 연관성이 낮아서 의미성이 낮은(less meaningful) 어휘를 사용하는 비율이 높다.

김소정과 전문기는 초등학교 6학년 학생의 에세이 176개를 상, 중, 하의 세 개의 집단으로 분류하여 수준별로 텍스트의 특성이 어떻게 다르게 나타나는지를 Coh-Metrix 프로그램을 활용하여 분석하였다 [6]. 어휘 지표의 경우, 단어 수, 타입 토큰 비율, 단어 빈도, 단어의 구체성, 이미지 형성 가능성, 어휘 습득 연령의 6개를 살펴보았다. 세 개의 집단 간의 차이점을 분석하기 위해 일원분산분석을 수행했으며, 단어 수를 제외하고는 상, 중, 하의 세 개의 집단 간에 유의미한 차이가 나타나지 않았다. 상 수준의 초등학교 6학년 학생은 많은 수의 단어를 사용하여 에세이를 작성했으나, 어휘가 다양하거나 정교화된 특성이 나타나지 않았다. 김소정과 전문기는 이러한 연구결과가 초등학교의 경우 영어 학습이 주로 학교에서 동일한 영어 교과서를 통해 이루어지고 영어 학습 시간도 길지 않기 때문으로 보았다. 즉, 상위집단의 학생이 하위 집단의 학생보다 어휘량은 높으나 다양성이나 정교화 지표에서는 통계적인 차이가 나지 않는 것으로 추정했다 [6].

어휘의 정교화를 살펴본 국내의 선행연구는 Coh-Metrix 프로그램을 활용하여 분석했으나, Coh-Metrix 프로그램은 어휘의 범위, n-gram, 학문 어휘 리스트와 같은 다양한 어휘지표를 분석할 수 없기 때문에 본 연구에서는 TAALES 프로그램을 활용하여 한국인 대학생이 작성한 에세이를 분석하고자 한다.

Ⅲ. 연구방법

3.1 연구대상

본 연구에서는 ICNALE(International Corpus Network of Asian Learners of English) 코퍼스에서 추출한 한국인 대학생이 작성한 에세이 600개를 분석하였다. 300명의 학생이 파트 타임 직업과 흡연 금지 주제에 대한 에세이를 각각 2개씩 작성했으며, 유럽공동 표준기준인 CEFR(Common European Framework of Reference)에 의하면 75명은 하 수준인 A2, 149명은 중 수준인 B1, 76명은 상 수준인 B2에 속한다.

본 연구에서 CEFR의 A2 집단은 하 수준, B1 집단은 중 수준, B2 집단은 상 수준으로 영어 능력 수준을 분류하여 분석한다.

3.2 어휘의 정교화 지표 추출

본 연구에서는 TAALES 프로그램이 생성한 어휘의 정교화 지표 중에서 정규분포와 다중공선성을 확인한 후 어휘의 정교화 지표를 추출하였다 [1]. 정규분포 가정에 위배되거나 다중공선성이 높아 어휘의 정교화 지표 간의 상관계수가 높게 나타난 지표를 삭제하였으며, 24개의 어휘의 정교화 지표가 판별분석에 투입되었다.

3.3 자료분석

본 연구에서는 300명의 한국인 대학생을 상, 중, 하의 세 개의 영어 수준으로 분류할 수 있는 어휘의 정교화 특성이 무엇인지를 살펴보기 위해 통계분석 프로그램인 SPSS 26.0을 활용하여 단계적 판별분석(a stepwise discriminant function analysis)이 수행되었다. 단계적 판별분석을 통해 통계적으로 유의미한 집단분류 능력이 있는 어휘의 정교화 지표에는 무엇이 있는지를 탐색하였다.

Ⅳ. 연구결과

본 연구에서 살펴본 연구질문은 300명의 한국인 대학생을 상, 중, 하의 세 개의 영어 수준으로 분류할 수 있는 어휘의 정교화 특성이 무엇인지를 살펴보는 것이며, 이를 위해 단계적 판별분석이 수행되었고, 결과는 <표 1>과 같다. 산출된 두 개의 함수는 통계적으로 유의미했다(함수 1 lambda=0.731, chi-square=186.093, p< 0.01; 함수 2 lambda=0.956, chi-square=25.438, p<0.01).

표 1. 판별분석 요약
Table 1. A summary of discriminant function analysis

	판별함수	함수 1	함수 2
구조행렬	SUBTLEXUS 코퍼스의 내용어 빈도	-0.664*	0.031
	내용어의 어휘 습득 연령	0.472*	0.142
	기능어의 어휘 결정 반응 평균 시간	0.447*	0.134
	상위어 동사	0.429*	-0.102
	COCA 학문 코퍼스의 기능어 빈도	0.203	0.613*
	기능어의 친숙도	-0.054	-0.522*
집단 중심점	하 수준 (A2 집단)	-0.745	0.229
	중 수준 (B1 집단)	-0.043	-0.210
	상 수준 (B2 집단)	0.818	0.185
정준상관계수		0.487	0.205
아이겐 값		0.311	0.044
설명변량 (%)		86.6	13.4
람다값		0.731	0.956

<표 1>에 제시된 집단중심점을 살펴보면 판별함수 1에 대해 하 수준인 A2 집단은 -0.745, 중 수준인 B1 집단은 -0.043, 상 수준인 B2 집단은 0.818이다. 따라서 판별함수 1은 상 수준을 하 수준과 중 수준으로 부터 구분하는 함수이다. 판별함수 2는 판별함수 1이 구분하고 난 나머지 집단을 구분하며 [7], 판별함수 2의 집단 중심점은 하 수준인 A2 집단은 0.229, 중 수준인 B1 집단은 -0.210으로 판별함수 2는 하 수준과 중 수준을 구분한다.

<표 1>의 설명변량에 제시된 것처럼 첫 번째 함수는 86.6%의 변량을 설명하며, 두 번째 함수는 13.4%의 변량을 설명한다. 두 개의 판별함수는 예측력이 있는 좋은 함수이며, 함수 1이 집단구분의 대부분인 86.6%를 차지한다. 판별함수 1은 판별함수 2보다 설명력이 높기 때문에 어휘의 정교화 지표에 의한 영어 수준 집단을 구별함에 있어 하 수준과 중 수준간의 차이보다는 상 수준을 중 수준과 하 수준으로부터 구별하는 차이가 더 크다고 볼 수 있다.

구조행렬은 판별함수와 예측변인간의 상관으로 판별함수에 대한 기여도를 나타낸다 [7]. 첫 번째 판별함수에 의해 상 수준을 하 수준과 중 수준으로 부터 구별 가능한 어휘의 정교화 지표에는 4개가 있으며, SUBTLEXUS 코퍼스의 내용어 빈도(SUBTLEXUS frequency content words, 구조행렬 계수= -0.664), 내용어의 어휘 습득 연령(age of acquisition content words, 구조행렬 계수=0.472), 기능어의 어휘 결정 반응 평균 시간 (lexical decision mean reaction time function words, 구조행렬 계수=0.447), 상위어 동사 (hypernymy verbs, 구조행렬 계수=0.429)로 나타났다.

판별함수 1에 의해 에세이를 유의미하게 구분하는 네 개의 어휘의 정교화지표의 평균값은 <표 2>에 제시되어 있다.

표 2. 어휘의 정교화 지표 평균 비교
Table 2. A comparison of mean of lexical sophistication features

	하 수준	중 수준	상 수준
SUBTLEXUS 코퍼스의 내용어 빈도	5.61	5.56	5.51
내용어의 어휘 습득 연령	4.38	4.42	4.48
기능어의 어휘 결정 반응 평균 시간	0.976	0.977	0.979
상위어 동사	1.57	1.63	1.70

SUBTLEXUS 코퍼스의 내용어 빈도의 경우 구조행렬 계수는 -0.664이며, 상 수준 학생의 평균은 5.51, 중 수준 학생의 평균은 5.56, 하 수준 학생의 평균은 5.61로 상 수준 학생의 평균이 가장 낮게 나타났다. 즉, 상 수준 학생의 경우 SUBTLEXUS 코퍼스에 빈번하게 나오는 어휘는 많이 사용하지 않았음을 나타낸다. 영어 능숙도가 높아질수록 빈번하게 출현하는 단어는 자주

사용하지 않게 된다 [5].

내용어의 어휘 습득 연령의 경우 구조행렬 계수는 0.472이며, 상 수준 학생의 평균은 4.48, 중 수준 학생의 평균은 4.42, 하 수준 학생의 평균은 4.38로 하 수준 학생의 평균이 가장 낮게 나타났다. 즉, 하 수준 학생은 어휘 습득 연령이 낮은 덜 정교화된 어휘를 많이 사용했으며, 상 수준 학생은 어휘 습득 연령이 높은 정교화된 어휘를 많이 사용했음을 알 수 있다.

기능어의 어휘 결정 반응 평균 시간의 경우 구조행렬 계수는 0.447이며, 상 수준 학생의 평균은 0.979, 중 수준 학생의 평균은 0.977, 하 수준 학생의 평균은 0.976으로 상 수준 학생의 평균이 근소한 차이로 높게 나타났다. 어휘 결정 과업에서 평균 반응시간이 길게 나타난 어휘는 정교화된 어휘를 말하며 상 수준 학생의 경우 근소한 차이로 정교화된 어휘를 많이 사용했음을 알 수 있다.

상위어 동사의 경우 구조행렬 계수는 0.429이며, 상 수준 학생의 평균은 1.70, 중 수준 학생의 평균은 1.63, 하 수준 학생의 평균은 1.57로 상 수준 학생의 평균이 가장 높고 하 수준 학생의 평균이 가장 낮게 나타났다. 즉, 상 수준 학생의 경우 상위어 점수가 높은 동사를 많이 사용한 특징을 나타낸다. 영어 능숙도가 높아질수록 구체적인 동사를 많이 사용하게 된다 [8].

<표 1>의 두 번째 판별함수에 의해 하 수준과 중 수준을 구별하는 어휘의 정교화 지표에는 2개가 있으며, COCA 학문 코퍼스의 기능어 빈도(COCA academic frequency function words, 구조행렬 계수=0.613), 기능어의 친숙도(familiarity function words, 구조행렬 계수=-0.522)로 나타났다.

판별함수 2에 의해 하 수준과 중 수준의 에세이를 유의미하게 구분하는 두 개의 어휘의 정교화 지표의 평균값은 <표 3>에 제시되어 있다.

표 3. 두 번째 판별함수에 의해 도출된 어휘의 정교화 지표 평균 비교
Table 3. A comparison of mean of lexical sophistication features based on the second discriminant function

	하 수준	중 수준
COCA 학문 코퍼스의 기능어 빈도	1.325	1.311
기능어의 친숙도	406.038	410.238

COCA 학문 코퍼스의 기능어 빈도의 경우 구조행렬 계수는 0.613이며, 중 수준 학생의 평균은 1.311, 하 수준 학생의 평균은 1.325로 하 수준 학생의 평균이 높게 나타났다. 이는 중 수준의 경우 COCA 학문 코퍼스에 빈번하게 나오는 단어는 자주 사용하지 않음을 시사한다. 영어 능숙도가 높아질수록 자주 등장하는 어휘는 사용하지 않게 된다 [9, 10].

기능어의 친숙도의 경우 구조행렬 계수는 -0.522로 나타났으며, 중 수준 학생의 평균은 410.238, 하 수준 학생의 평균은 406.038로 중 수준 학생의 평균이 더 높게 나타났다. 영어 능숙도가 높아질수록 친숙한 단어를 덜 사용하게 되는데 [5], 본 연구결과는 선행연구와 상반된다. <표 2>에서 살펴본 것처럼 하 수준에서 빈도가 높은 내용어를 더 많이 사용한 것으로 나타났는데, 기능어의 친숙도의 경우 왜 중 수준에서 하 수준보다 더 자주 사용했는가에 대해서는 후속연구가 더 필요하다.

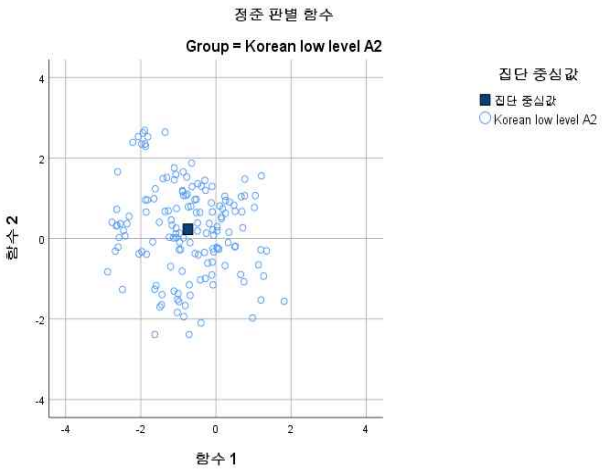

그림 1. 하 수준 학생의 판별함수
Figure 1. A discriminant function for A2 level

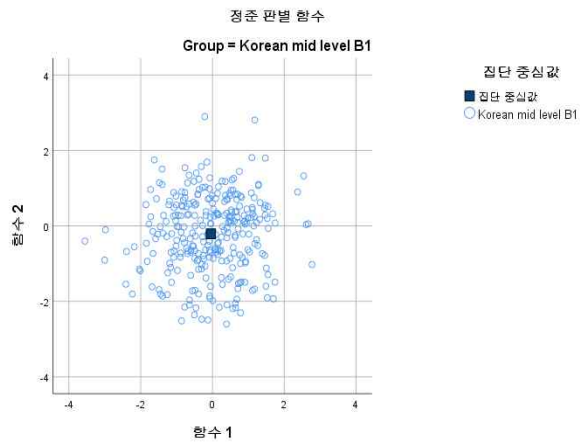

그림 2. 중 수준 학생의 판별함수
Figure 2. A discriminant function for B1 level

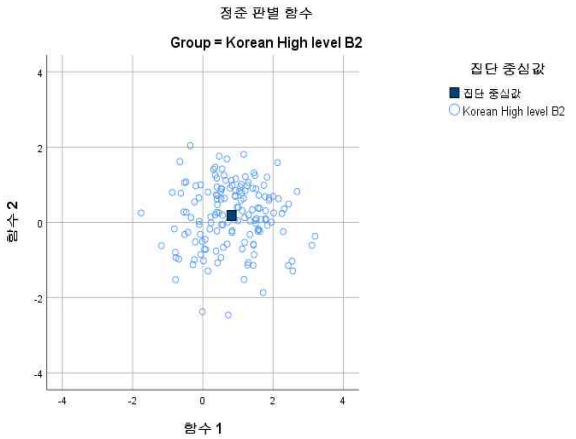

그림 3. 상 수준 학생의 판별함수
Figure 3. A discriminant function for B2 level

판별함수 1을 CEFR의 A2, B1, B2의 영어수준별로 도식화 하면 <그림 1>, <그림 2>, <그림 3>과 같다. 하 수준인 A2의 경우 판별함수 평균은 -0.745, 중 수준인 B1의 판별함수 평균은 -0.043, 상 수준인 B2의 판별함수의 평균은 0.818이고, <표 1>에 제시된 집단 중심점과 같다. 상 수준 학생의 집단 중심점이 가장 높았는데 이는 상 수준 학생의 경우 전반적으로 어휘의 정교화 지표를 다른 집단의 학생에 비해 많이 사용했음을 단적으로 보여준다.

전체 학생의 판별함수는 <그림 4>와 같다. 판별함수 X축을 기준으로 상 수준은 오른쪽에, 중 수준은 가운데에, 하 수준은 왼쪽에 분포되어 있다.

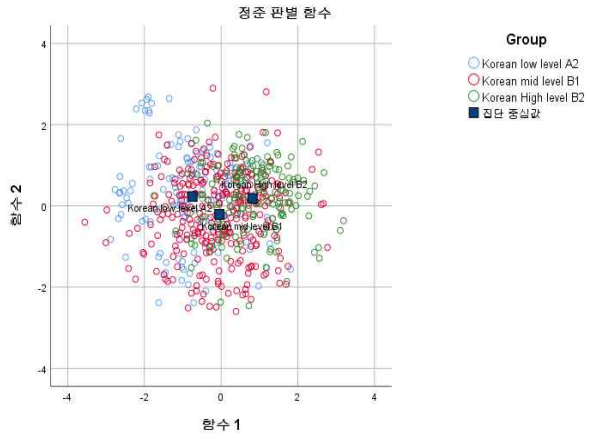

그림 4. 전체 학생의 판별함수
Figure 4. A discriminant function for A2, B1, B2 level

도출된 판별함수에 의한 하 수준, 중 수준, 상 수준 집단의 분류 결과는 <표 4>에 제시되어 있다.

표 4. 판별함수에 의한 분류 결과
Table 4. Classification result by the discriminant function

	예측된 하 수준 에세이	예측된 중 수준 에세이	예측된 상 수준 에세이	전체
실제 하 수준 에세이	86 (57.3%)	38 (25.3%)	26 (17.3%)	150 (100%)
실제 중 수준 에세이	96 (32.2%)	106 (35.6%)	96 (32.2%)	298 (100%)
실제 상 수준 에세이	16 (10.5%)	29 (19.1%)	107 (70.4%)	152 (100%)

하 수준인 A2의 경우 전체 150개의 에세이 중에서 86개를 A2 수준으로 정확하게 분류했으며, 중 수준인 B1의 경우 전체 298개의 에세이 중에서 106개를 B1 수준으로 정확하게 분류했고, 상 수준인 B2의 경우 전체 152개의 에세이 중에서 107개를 B2 수준으로 정확하게 분류하는 것으로 나타났다. 즉, 전체적으로 정확하게 분류한 비율은 49.8%이고 상 수준인 B2는 70.4%를 정확히 분류한 반면에, 하 수준인 A2는 57.3%로 나타났고 중 수준인 B1을 정확히 분류한 비율은 35.6%로 가장 낮게 나타났다.

요약하면, 판별분석에서 도출된 6개의 지표는 상 수준과 하 수준 학생의 에세이를 정확하게 분류하는 비율은 높았으나, 중 수준의 경우는 상대적으로 정확도가 떨어진다고 볼 수 있다. 이는 앞에서 설명한 것처럼 판별함수 1의 설명력이 높고 집단구분의 대부분을 차지하며 판별함수 2는 설명력이 낮았던 것과 연관된다.

V. 결 론

본 연구에서는 TAALES 프로그램을 활용하여 한국 대학생을 상, 중, 하의 세 개의 영어 수준으로 분류할 수 있는 어휘의 정교화 특성이 무엇인지를 살펴보았다. 본 연구에서는 SPSS 26.0의 프로그램을 활용하여 판별분석이 수행되었다.

연구분석 결과를 요약하면 다음과 같다. 판별분석 결과 첫 번째 판별함수에 의해 상 수준을 하 수준과 중 수준으로부터 구별 가능한 어휘의 정교화 지표에는 4개가 있으며, SUBTLEXUS 코퍼스의 내용어 빈도,

내용어의 어휘 습득 연령, 기능어의 어휘 결정 반응 평균 시간, 상위어 동사로 나타났다. 상 수준 학생의 경우 SUBTLEXUS 코퍼스에 자주 나오는 어휘는 많이 사용하지 않았고, 어휘 습득 연령이 높은 고급 수준의 정교화된 어휘를 많이 사용했음을 알 수 있다. 또한 상 수준 학생은 어휘 결정 과업에서 평균 반응시간이 길게 나타난 정교화된 어휘를 많이 사용했으며, 상위어 점수가 높은 구체적인 동사를 많이 사용한 특징이 있다.

판별분석에서 도출된 어휘의 정교화 지표는 상 수준과 하 수준 학생의 에세이를 정확하게 분류하는 비율이 각각 70.4%, 57.3%로 높게 나타났다. 본 연구에서 도출된 첫 번째 판별함수가 변량의 대부분인 87.7%를 차지하기 때문에, 두 번째 판별함수에 의해 중 수준과 하 수준을 구별하는 어휘의 정교화 지표는 정확도가 떨어지며 특히 중 수준의 경우 분류 정확도가 35.6%로 상대적으로 낮게 나타났다.

References

[1] Kyle, K., & Crossley, S. Automatically assessing lexical sophistication: Indices, tools, findings, and applications. TESOL Quarterly, 49(4), pp. 757-786. 2015. DOI : 10.1002/tesq.194|

[2] Crossley, S. A., & McNamara, D. S. Computational assessment of lexical differences in L1 and L2 writing. Journal of Second Language Writing, 18, pp. 119-135. 2009. DOI : 10.1016/j.jslw.2009.02.002

[3] Crossley, S. A., & McNamara, D. S. Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. Journal of Research in Reading, 35, pp. 115-135. 2012. DOI : 10.1111/j.1467-9817.2010.01449.x|

[4] Guo, L., Crossley, S. A., McNamara, D. S. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. Assessing Writing, 18, pp. 218-238. 2013. DOI : 10.1016/j.jslw.2013.05.002

[5] Jung, Y., Crossley, S. A., & McNamara, D. S. Linguistic Features in MELAB Writing Task Performances. CaMLA Working Papers. 2015.

[6] Kim, Sojung, & Jeon, Moongee. An analysis study of English writing of elementary school 6th grade English language learners using Coh-Metrix. Modern English Education, 17(3), 263-287. 2016.

[7] Yang, Byug-Hwa. Understanding Multivariate Data Analysis. Seoul: Communication Books. 2006.

[8] Kyle, K., & Crossley, S. The relationship between lexical sophistication and independent and source-based writing. Journal of Second Language Writing, 34, pp. 12-24. 2016. DOI : 10.1016/j.jslw.2016.10.003

[9] Laufer, B., & Nation, P. Vocabulary size and use: Lexical richness in L2 written production. Applied Linguistics, 16, pp. 307-322. 1995. DOI : 10.1093/applin/16.3.307

[10]Nation, P. Learning vocabulary in another language. Cambridge, England: Cambridge University Press. 2013.