MALL

Multimedia-Assisted Language Learning
23(2) 215-235

# 한국인 영어 학습자의 쓰기능력별 어휘 및 연어 사용 양상 분석

**김성연**(한양대학교)
**신동광**(광주교육대학교)
**김경숙**(한양대학교)

Kim, Sung Yeon, Shin, Dongkwang & Kim, Kyung-Sook (2020). Korean college students' use of vocabulary and collocation according to writing proficiency. *Multimedia-Assisted Language Learning, 23*(2), 215-235.

This study examined both lexical and collocational knowledge of Korean college students in relation to their writing proficiency measured by a diagnostic writing test. The study analyzed essays written in response to two topics by college students at three different proficiency levels (high, mid, low). In a comparison of vocabulary used in the essays, the study noted differences between the proficiency levels in terms of token and type frequencies. However, the lexical distribution patterns in reference to graded word-lists were similar across the proficiency levels. Regardless of the topic, approximately 90% of words used in the essays were from the first 2K words, and about 95% of words from the first 3K words. As to learner use of collocation, collocational expressions were more frequently used by advanced learners, compared to intermediate- or low-level learners. Despite the difference, collocational distribution patterns according to the graded collocation-list were also similar across the three proficiency levels. Regarding part-of-speech-tagged collocation, students tended to overuse a limited set of patterns, such as AJ-NN, NN-NN, and VP-AVP. Findings are discussed in more detail, along with pedagogical implications.

**Key words** corpus analysis, vocabulary, productive vocabulary, collocation, writing proficiency, part of speech tagging
*doi: 10. 15702/mall.2020.23.2.215*

## I. 서 론

언어학습에서 어휘 지식의 중요성은 많은 연구들을 통해 강조되었다(Laufer & Hulstijin, 2001; Lewis, 1993; Milton, 2013; Nation, 2001, 2006; Schmitt, 2008). 어휘를 조합하여 구문을 만들고, 구문을

담화적으로 구성하여 텍스트를 산출하기 때문에 텍스트의 기본 단위로서 어휘의 역할은 중요하다(Ghadessy, 1989). 어휘는 듣기, 말하기, 읽기, 쓰기 기능의 근간을 구성하는 요소로서 다양한 방식(명시적, 암시적, 직접적, 간접적, 독립적, 통합적 방법)으로 교수되어 왔다(Laufer & Hulstjin, 2001; Nation, 2001; Paribakht & Wesche, 1997; Schmitt, Cobb, Horst, & Schmitt, 2017).

어휘에 대한 관심은 많은 연구들로 이어졌는데 컴퓨터 기술의 발달과 함께 언어 데이터베이스인 말뭉치(corpus) 구축이 가능해지고 콘코던스(concordancer)를 포함하여 다양한 분석용 프로그램이 개발됨에 따라 방대한 양의 어휘 연구들이 발표되었다(Cobb, 2010; Read, 2000). 또한, British National Corpus(BNC), Corpus of Contemporary American English(COCA), 또는 이를 통합한 BNC-COCA 등의 말뭉치로부터 많은 어휘목록(word list)들이 추출됨에 따라 이를 준거로 하여 학습자 언어의 특성을 대조, 비교하는 연구들이 많이 이루어졌다(Dang & Webb, 2014; Gregori-Signes & Clavel-Arroita, 2015; Kao & Wang, 2014; Kim & Ryu, 2009, 2011; Laufer & Nation, 2001; Nation, 2006; Paquot, 2007; Shin, 2015).

그러나 기존에 이루어진 많은 연구들은 학습자가 구사하는 어휘가 등급화된 어휘목록별로 어떻게 분포되어 있는지 어휘 수(token)나 유형 수(type) 중심으로 기술하는 데 국한되어 있다(Imura, 2002; Leech, Rayson, & Wilson, 2001; Shimamoto, 2000). 문제는 이와 같은 방식이 학생들의 어휘 지식 수준을 정확하게 분석·파악하기에는 제한적이라는 점이다. 특히, 문어 담화의 52.3%가 숙어, 연어 등의 정형화된 어구(formulaic sequence)로 구성되어 있음을 감안했을 때 학습자의 어휘 지식을 분석할 때 개별 어휘와 함께 연어에 대한 지식을 함께 파악하는 것이 중요하다(Ehrman & Warren, 2000). 어휘 지식은 단어의 의미와 형태의 관계를 아는 것 이상을 의미하기 때문에 개별 단어뿐 아니라 연어의 의미와 사용을 아는 것이 중요하다(Tsai, 2015). Bahns와 Eldaw(1993)는 개별 단어에 대한 학습자 지식만으로는 어휘 지식을 설명할 수 없다고 지적하며 개별 어휘뿐 아니라 연어를 아는 것이 중요하다고 주장하였다. Lewis(1993)의 'Lexical Approach'를 계기로 연어와 같은 '조립식 문형(prefabricated patterns)'에 대한 관심이 높아졌는데 그 결과 연어에 대한 연구도 다수 이루어졌다(Altenberg & Granger, 2001; Cross & Papp, 2008; Durrant & Schmitt, 2009; Siyanova-Chanturia, 2015).

문제는 대부분의 선행연구들이 어휘와 연어를 분리하여 분석·보고하고 있다는 점이다. 이를 동시에 분석한 연구들이 있지만(신동광, 전유아, 이신웅, 박명수, 2018; Bestgen, 2017; Vedder & Benigno, 2016), 다어휘 표현 분석이 두 개의 단어 조합에 국한되어 이루어졌거나, 표집의 크기가 작거나, 빈도분석이 최상위 빈도의 어휘와 다어휘 표현 중심으로 이루어졌다는 점에서 제한적이다. 예를 들어, 신동광 외(2018)는 EFL 학습자 코퍼스와 원어민 코퍼스의 개별어휘 및 다어휘 표현 빈도와 함께 최상위 20개 어휘와 다어휘 표현을 기술하는 데 초점을 두었다. Vedder와 Benigno(2016)는 39명의 중하위권과 중위권 수준의 학습자가 쓴 글을 표집하여 분석했다는 점, Bestgen(2017)은 글감의 질을 결정하는 요인을 조사하기 위해 개별 어휘(single-word)와의 비교에 두 개 단어 조합(bi-gram)만 고려했다는 점에서 한계가 있다.

또한 학습자의 쓰기능력 수준에 따라 어휘와 연어를 동시에 비교한 연구도 찾아보기 힘들다.

Siyanova-Chanturia(2015)는 코퍼스 연구 다수가 중상위 이상 수준의 학습자에 편중되어 분석이 이루어진 것을 한계점으로 지적하였다. 상·중·하 수준별로 학습자의 어휘와 연어 사용을 분석하는 것은 학문적으로 의의가 있고 학습자 어휘/연어 수준에 대한 현장교사의 이해를 도모한다는 점에서 실용적 가치가 있다. 특히, 문어담화에서의 어휘 및 연어 사용을 조사하고자 할 때 쓰기능력 수준별로 분석하는 것이 타당하다. 즉, 어휘/연어 평가, C-test 등의 간접평가보다는 직접평가인 에세이 쓰기평가 결과에 따라 수준을 나누어 에세이에 사용된 표현어휘와 연어를 수준별로 살펴보는 연구가 필요하다. 또한 연어의 경우 많은 선행연구들이 연어 패턴을 총체적으로 분석하지 않고 형용사-명사, 동사-명사 등의 특정 패턴에 집중했다는 점에서 한계가 있다(Durrant & Schmitt, 2009; Kashiha & Chan, 2015; Laufer & Waldman, 2011; Li & Schmitt, 2010; Siyanova-Chanturia, 2015). 이에 본 연구는 대학 신입생들의 영어 쓰기평가 결과를 기준으로 쓰기 수준을 구분하고 개별 어휘 및 연어와 같은 다어휘 표현(multi-word units)의 사용 특성을 수준별로 비교, 분석하고자 한다.

# II. 이론적 배경

## 1. 어휘 지식과 쓰기능력의 관계 연구

어휘는 언어 입력의 이해와 출력의 산출과 관련된 텍스트를 구성하는 기초 단위로서 듣기, 말하기, 읽기, 쓰기 등의 기능 습득에 직·간접적인 영향을 준다(Stæhr, 2008). 어휘를 안다는 것은 입력에 나타난 단어를 인지하는 이해어휘(Receptive Vocabulary, RV)지식뿐 아니라 출력산출을 위해 사용하는 표현어휘(Productive Vocabulary, PV)지식까지 포함한다. 어휘연구의 동향을 보면 읽기 등의 언어 입력에서 어휘를 인지할 수 있는 능력, 즉 이해어휘 지식에 집중한데 반해 표현어휘에 대한 관심은 상대적으로 낮은 것을 알 수 있다. Laufer와 Nation(1999)의 표현어휘 평가(Productive Vocabulary Level Test, PVLT) 외에는 검사 도구가 많지 않아서 C-test 등이 사용되기도 했다. 그러나 C-test 같이 통제된 표현어휘 평가는 어휘 사용 능력을 직접적으로 측정하기에 제한적이어서 최근에는 쓰기에 나타난 어휘 사용을 직접적으로 관찰함으로써 어휘 지식과 쓰기능력의 관계를 설명하려는 연구들이 이루어졌다(신동광, 2018). 원어민 기준으로 개발된 어휘 검사지는 EFL 환경의 사회문화적 변인(예, 외래어, 입시위주의 어휘교육)을 반영하지 못하여 제한적이기 때문이다(Lee, Chon, & Shin, 2012).

Douglas(2013)는 캐나다의 한 대학에서 학생들의 표현어휘 특성을 분석하기 위해 쓰기평가를 실시한 후, 시험을 본 12개 전공의 중·상급 수준(Marginally Satisfactory or Satisfactory)의 대학 신입생 120명(남학생 64명, 여학생 56명)의 답안으로 학습자 코퍼스를 구축하였다. 학생들은 2시간 30분 동안 진행된 쓰기평가에서 4개의 학술 주제 중 하나를 선택하여 400단어 내외의 설명문(expository essay)을 작성하였고 이를 바탕으로 총 62,309 단어의 학습자 코퍼스가 구축되었다. 이 학습자 코퍼스의 어휘를 분석한 결과, 최상위 2,000 단어로 구성된 어휘목록 GSL(West, 1953)이 커버하는 범위는 약 87.65%였고 570단어로 구성된 학술어휘목록 AWL(Coxhead, 2000)이 차지하는 비중은 6.74%인 것으로 나타났다.

학술적 주제에 대한 글쓰기였기 때문에 학술 어휘의 비중은 일반 에세이(4% 미만)보다 높게 나타났다. Cobb(2012)의 BNC-20을 적용하여 분석했을 때는 상위빈도 3,000 단어가 학습자 코퍼스를 구성하는 어휘의 약 95%를 커버했고 98%까지 커버하기 위해서는 5,000 단어의 지식이 필요한 것으로 나타났다. 기존의 연구들이(Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt, Jiang, & Grabe, 2011) 듣기, 읽기 영역에서 대학생에게 요구되는 어휘 지식을 5,000 단어 수준으로 보고한 바 있는데 Douglas(2013)는 쓰기 영역에서도 5,000 단어 이상의 어휘 지식이 필요하다고 제안하였다. Douglas는 복수채점을 통해 학습자 에세이를 상·중·하 3개의 등급으로 구분하였으나 수준별로 어휘 분석을 하지 않았다는 점에서 한계가 있다.

쓰기능력을 예측하기 위한 도구로서 어휘 분석을 한 연구도 있는데, Yüksel(2015)은 40명의 터키 대학생들을 대상으로 어휘 수준 평가를 실시하였다. 어휘 평가를 위해 Nation(1990)의 어휘평가의 5개 등급(2nd 1000, 3rd 1000, 5th 1000, 10th 1000, 학술어휘)에서 각 30개 문항을 추출하여 검사지를 구성하였다. Yüksel은 어휘 지식과 쓰기능력 간의 관계를 보기 위해 학생들에게 TOEFL 쓰기 주제를 주고 30분간 300단어 정도의 에세이를 작성하게 한 후 GSL과 AWL 어휘목록을 기준으로 실제 사용된 표현어휘 수준을 측정하였다. 분석 결과 학생들의 어휘 점수는 3rd 1000 단어 수준에서 30개 문항 중 평균 16.58점으로 나타났는데, 2nd 1000 단어 수준(평균 16.25)보다 높게 나와서 원어민 기준으로 개발된 어휘 검사지가 EFL 환경의 사회문화적 특성을 반영하는 데 한계가 있음을 보여주었다. 한편, 학생들이 에세이 쓰기에 사용한 표현어휘는 2000 단어로 구성된 GSL이 커버하는 비중은 92.68%로 나타났고 AWL은 5.42%로 나타났다. 에세이 쓰기에 사용된 표현어휘 지식과 쓰기능력의 관계는 유의하지 않았지만 어휘 평가를 통해 측정한 이해어휘 지식과 쓰기능력은 유의한 상관관계를 보였다. 단순회귀분석을 통해 쓰기능력에 대한 어휘 지식의 예측도를 분석한 결과에서도 이해어휘 지식이 쓰기능력을 유의한 수준에서 설명하는 것으로 나타났다($F = 16.315$, $p < .05$).

최근 Kiliç(2019)은 54명의 터키 대학생들을 대상으로 이해어휘 평가(Receptive Vocabulary Levels Test, RVLT, Schmitt, Schmitt & Clapham, 2001)를 실시하여 이해어휘 지식을 측정하였다. 또한, 표현어휘 평가(PVLT, Laufer & Nation, 1999)형식을 참고하여 이해어휘 검사지(RVLT)를 빈칸 채우기 유형으로 변환한 평가를 개발하였다. RVLT와 PVLT를 활용하여 2nd 1000, 3rd 1000, 5th 1000, 10th 1000, UWL 수준을 측정하였고, 학생들이 작성한 논설 에세이에 대해 과제성취, 구성, 영어 사용, 어휘, 맞춤법 및 철자 등의 영역으로 구분하여 4점 척도로 채점하였다. 분석 결과, 이해어휘와 표현어휘의 평균 점수는 각각 40.5%, 38.6%의 정답률을 보여 표현어휘 지식이 이해어휘 지식보다 낮은 것으로 나타났다. 이해어휘와 표현어휘의 상관계수는 $p < .001$ 수준에서 .87로 매우 높게 나타났다. 이해어휘, 표현어휘 점수와 쓰기 점수의 상관계수는 $p < .001$ 수준에서 각각 .49와 .48로 나타났다. 그러나 다중회귀분석 결과에서는 표현어휘 지식이 이해어휘 지식보다 쓰기능력을 더 잘 예측하는 것으로 나타났다. 이상의 연구들을 통해 쓰기능력과 어휘 간의 관계를 볼 수 있으며 쓰기에 사용된 어휘는 표현어휘이기 때문에 표현어휘의 중요성을 알 수 있다.

## 2. 연어 지식과 쓰기능력의 관계 연구

어휘와 달리 연어의 경우에는 연어 사용 수준을 측정하는 등급별 검사지가 많지 않고 그 결과 연어 지식의 절대량을 측정, 검사한 연구도 찾아보기 힘들다(Martinez & Schmitt, 2012). 연어 지식을 측정하기 위해 학술어휘 목록(UWL)에서 추출한 단어의 연계어로 구성된 WAT (Word Associate Test, Read, 1993)나 COLLEX/COLLMATCH(Gyllstad, 2009) 등의 연어 검사지를 적용하기는 하지만 이 또한 절대적인 연어 지식의 양을 측정하기보다는 측정 방식에 초점을 두고 있다. 때문에 연어 지식과 쓰기능력의 관계에 대한 연구는 연어 지도가 쓰기능력에 미치는 영향이나 쓰기 답안에 나타난 연어 사용 패턴 등에 집중하는 연구들이 주류를 이루고 있다.

예를 들어, Siyanova-Chanturia(2015)는 이탈리아어를 배우는 중국인 학습자 36명의 글에 사용된 명사-형용사 연어 패턴이 학기 초반, 중반, 후반에 어떻게 변화하는지 분석하였다. 그 결과 초반보다 후반에 작성된 글에서 구성 성분 간의 조합(association)이 더 강하고 빈도 높은 연어들이 많이 사용되는 것을 발견하였다. 한편, 연어 검색 도구의 효과를 비교한 Nurmukhamedov (2017)는 미국 대학의 ESL 프로그램에 등록된 45명의 성인들을 세 집단으로 나누어 Longman Dictionary of Contemporary English (LDOCE)<https://www.ldoceonline.com> 온라인판, Macmillan Collocation Dictionary(MCD) 인쇄본, WordNet의 사전적 정의를 제공하고 연어 검색을 지원하는 WordandPhrase.Info(WPI, www.wordandphrase.info)를 도구로 제공하였다. 학생들은 각각의 도구 활용 방법을 익힌 후에 에세이에 포함된 16개의 부자연스런 연어(8개의 형용사-명사, 8개의 동사-명사 조합)표현에서 명사 외의 부분(형용사와 동사)을 수정하였는데, 집단 간에 도구 활용 순서를 교차하는 방식을 통해 모든 학생이 세 가지 도구를 활용할 수 있게 하였다. 도구별로 학생들의 과업수행을 비교한 결과 정답률이 LDOCE(10.69점), WPI(10.58), MCD(8.44) 순으로 나타났다. 이와 같은 차이는 반복측정 변량분석에서도 통계적으로 유의한 것으로 확인되었다. 한편, 학생들의 태도 조사에서는 WPI에 대한 긍정적 반응이 우세한 것으로 나타났다. 학생들은 연어의 구성소(collocates)를 많은 예시와 함께 제공하는 WPI가 연어 검색 및 수정에 유용하다고 인식하고 있었다.

연어 연구의 다른 주류는 쓰기에 나타난 연어의 오류를 유형, 빈도 중심으로 분석한 연구들이다. 대표적인 예로 신동광, 배주경, 송민영(2014)은 국가영어능력평가시험 쓰기 문항에 대한 한국 고등학생들의 답안(n=1119)을 분석하여 학생들이 가장 빈번하게 범하는 연어 오류 유형을 조사하였다. 이를 위해 2급과 3급 평가의 문항 유형 두 가지에 대해 A(상), B(중), C(하) 등급에 해당하는 답안을 약 70～100개씩 추출하여 구성하였다. 답안의 길이는 2급의 경우 60～120 단어, 3급의 경우 40～50단어 내외였는데, 분석 결과 쓰기 문항 유형이나 쓰기능력 수준에 관계없이 '단어 선택 오류(예, The [middle>mid-term] exam was tough.)'가 가장 많았고 '전치사 선택 오류(예, I agree [about>with] joining the club activity.)'나 'L1-L2 번역 오류(예, I [did>got] the 3rd rank in the science quiz.)'가 뒤를 이었다. 주목할 점은 언어 오류 측면에서 상위권 학습자의 답안에 나타난 오류 수가 중하위권 학습자의 것보다 상대적으로 적었지만 연어 오류의 경우 상위권 학습자의 답안에서 더 많이 관찰되었다. 유사한 결과가 Shitu(2015)에서도 보고되었는데 300명의 나이지리아 ESL 대학생들이 쓴 900개의 에세이에 나타난 연

어 오류를 분석한 Shitu는 상위권 학생들도 모국어 간섭, 과잉일반화, 연어지식의 부족으로 인해 연어 관련 오류를 범한다고 설명하였다. Vedder와 Benigno(2016)는 연어가 늦은 시기에 습득된다고 주장하였고 Laufer와 Waldman(2011)은 연어의 발달이 더디고 불규칙적이라고 경고하였다. 실제 많은 선행연구들을 통해 L2 학습자가 연어를 어려워하며 능숙도나 학습 기간에 관계없이 오류를 범한다는 사실이 보고되었다(Kuo, 2009; Laufer & Waldman, 2011; Nesselhauf, 2005; Shitu, 2015; Vedder & Benigno, 2016).

그럼에도 불구하고 코퍼스 연구는 연어보다는 개별 어휘에 집중하여 이루어졌던 것이 사실이다. 이는 앞서 밝힌 바와 같이 이해어휘, 표현어휘를 측정하는 평가 도구의 개발, 코퍼스 구축, 콘코던서 등을 포함한 어휘 분석 도구의 개발로 인해 어휘 분석이 용이해졌기 때문이다(Coxhead, 2016). 다만 최근에는 연어, 조립식 문형 등을 포함한 다어휘에 대한 관심이 높아짐에 따라 관련 연구들도 점진적으로 증가하고 있는 추세이다. 그러나 연어를 체계적으로 분석할 수 있는 프로그램이 많지 않아 연구의 내용이나 범위 면에서 제한적이었다. 예를 들어 형용사-명사, 동사-명사 등과 같은 특정 품사 조합이나 특정 연어 표현에 집중한 연구들이 많이 이루어졌다(Kim & Le, 2018; Laufer & Waldman, 2011; Li & Schmitt, 2010; Siyanova-Chanturia, 2015; Siyanova-Chanturia & Schmitt, 2008). 예를 들어, Laufer와 Waldman(2011)은 이스라엘의 EFL 학습자 코퍼스를 구축하고 동사-명사 연어 패턴을 학습자 수준별로 비교하고 학습자와 원어민을 비교한 결과 원어민이 L2 학습자보다 동사-명사 패턴을 더 많이 사용하고 상수준의 학습자가 중, 하 수준의 학습자보다 해당 패턴을 많이 사용하는 것을 발견하였다. 그러나 이와 같이 특정 품사 패턴에 집중하는 경우 학생들의 연어 사용 패턴 전체를 조명하지 못하는 한계점이 있다. 이에 본 연구는 한국의 EFL 대학생들이 쓴 글에 나타난 연어의 품사 조합 전체를 분석하여 연어 사용을 종합적으로 살펴보고 학습자 수준별 연어 사용 양상을 기술하고자 한다. 연어와 함께 어휘 사용도 학습자 수준에 따라 살펴보고자 한다. 본 연구는 어휘 지식을 구성하는 어휘와 연어 두 가지 구성요소를 동시에 고찰하고, 그 사용 양상을 쓰기능력 수준별로 분석, 기술한다는 점에서 의미가 있다.

# III. 연구 방법

## 1. 연구 질문

앞서 기술한 연구목표를 질문으로 구체화하면 다음과 같다.

1) 한국인 영어 학습자의 영어 쓰기에 나타난 어휘 사용은 쓰기능력별로 어떤 양상을 보이는가?
2) 한국인 영어 학습자의 영어 쓰기에 나타난 연어 사용은 쓰기능력별로 어떤 양상을 보이는가?

## 2. 연구 자료

본 연구의 데이터는 서울에 소재한 대학에서 신입생을 대상으로 실시한 영어쓰기 진단평가 자료에 기초하고 있다. 연구를 위해 대학영어 프로그램을 주관하는 창의융합교육원에서 최근 5년간 실시한 진단평가 자료를 활용하여 영어학습자 문어 코퍼스(Hanyang English Learner Corpus, HELC)를 구축하였는데, 본 논문에서는 HELC 데이터의 2018년 자료 중 일부를 중심으로 분석, 보고하고자 한다.

2018년 진단평가는 온라인 평가시스템을 통해 두 개의 쓰기 문항(문항 A, 문항 B) 중 한 문항이 응시자에게 임의 배정되는 방식으로 진행되었다. 학생 답안의 총 수는 2,303개였으며, 문항 B에 대한 답안(n=1,162)이 문항 A에 대한 답안(n=1,141)보다 많았다. 대학에서 자체 개발한 채점 기준을 활용하여 원어민 교수들이 답안을 채점하였는데 그 점수 값을 참조하여 수준을 구분하였다. 문항 A와 B에 대해 세 개의 수준(상, 중, 하)에서 층화 추출(stratified sampling)을 통해 50명의 답안을 임의로 선정하는 방식으로 각 150명씩 총 300명의 답안을 추출하여 데이터를 구성하였다.

〔표 1〕 쓰기평가 문항

| | |
|---|---|
| 문항 A | Do you agree or disagree with the following statement? Television, newspapers, magazines, and other media pay too much attention to the personal lives of famous people such as public figures and celebrities. Use specific reasons and details to explain your opinion. |
| 문항 B | Do you agree or disagree with the following statement? Face-to-face communication is better than other types of communication, such as letters, E-mail, or telephone calls. Use specific reasons and details to support your answer. |

문항별 답안에 대한 기술통계는 [표 2]와 같은데 학생들의 수준이 높아짐에 따라 답안이 길어진 것을 알 수 있다. 한편, 주제별 답안 길이의 차이는 크지 않았는데, 문항 B에 대한 답안의 총 단어 수(token)는 49,106개로 문항 A에 대한 답안(46,190)보다 더 많았다. 그러나 문항 B에 대한 답안 수가 A에 대한 답안 수보다 21개 더 많았던 것을 감안하면 그 차이가 크지 않았음을 알 수 있다.

〔표 2〕 쓰기 답안 데이터 구성

| 문항 A | | | 문항 B | | |
|---|---|---|---|---|---|
| 등급 | 단어수 | 답안수 | 등급 | 단어수 | 답안수 |
| 상 | 20,102 | 50 | 상 | 20,483 | 50 |
| 중 | 15,995 | 50 | 중 | 16,464 | 50 |
| 하 | 10,093 | 50 | 하 | 12,159 | 50 |

| | | | | | |
|---|---|---|---|---|---|
| 계 | 46,190 | 150 | 계 | 49,106 | 150 |

## 3. 분석 도구

### 1) 어휘 분석 프로그램

어휘 분석을 위해 Nation(2012)이 개발한 BNC-COCA25 Range Program을 사용하였다. 이 프로그램에 탑재된 어휘목록 BNC-COCA25는 미국 코퍼스 Corpus of Contemporary American English(COCA)와 영국 코퍼스 British National Corpus(BNC)의 어휘를 통합하여 추출한 가장 대표성 있는 어휘목록으로 등급별로 1,000개 단어씩 최대 25개 등급(25,000 단어) 수준까지 분석할 수 있는 장점이 있다. 본 연구에서는 한국인의 최대 영어 어휘 지식량과 연어 분석 시 등급수를 고려하여 20개 등급에 해당하는 20,000 단어 수준까지만 분석하였다.

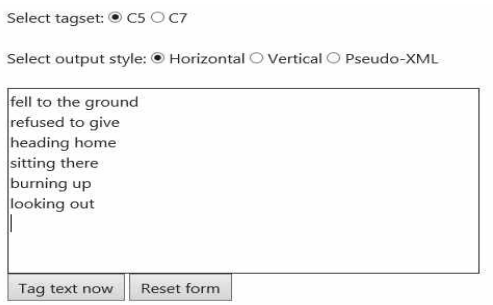
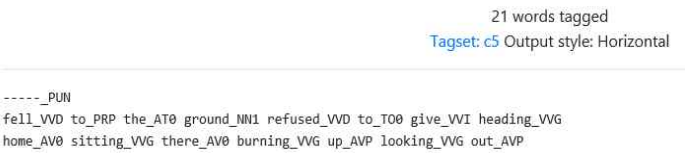
### 2) 연어 분석 프로그램

#### (1) COCA_MWU20 ColloGram

한국 대학생들이 영어 쓰기에서 어떤 유형과 수준의 연어를 사용하는지 조사하기 위해 연어 사용의 양과 종류를 등급별로 체계적으로 분석할 수 있는 COCA_MWU20 ColloGram(Shin, Chon, Lee, & Park, 2018)을 사용하였다. COCA_MWU20 ColloGram은 COCA에서 추출한 연어목록이 탑재된 프로그램으로 등급별로 500개 연어씩 최대 20개 등급(10,000개) 수준까지 분석할 수 있다. 연어 선별을 위해 1990년~2009년의 COCA 데이터(4억 5천만 단어)에서 최소 20회 이상의 빈도를 가지고 하나의 독립된 의미 단위를 구성하며 COCA의 5개 대영역 중 최소 4개 영역에서 출현하는 조건(Range 4 이상)을 적용하였다. COCA_MWU20 ColloGram은 정확히 반복된 단어 조합의 패턴만을 검색하는 기존의 연어 분석 프로그램과는 달리 연어 목록(COCA_MWU20)에 포함된 연어의 사용 양상을 등급별로 분석한다.

#### (2) CLAWS Web Tagger

본 연구는 쓰기에 사용된 연어의 종류뿐 아니라 연어 구성소(collocates)들의 품사 조합 패턴을 분석하였다. 품사 태깅(POS tagging)의 신뢰성을 확보하기 위해 학생들의 쓰기 답안을 COCA_MWU20 ColloGram을 통해 분석한 후 CLAWS Web Tagger(University Centre for Computer Corpus Research on Language, UCREL, 2014)를 활용하여 쓰기에 사용된 연어의 품사를 태깅하였다.

<table>
<tr><td>연어<br>입력</td><td> 
</td></tr>
<tr><td>연어<br>태깅<br>산출</td><td> 
</td></tr>
<tr><td>품사<br>태그<br>통합<br>(13개)</td><td>AJ: Adjective<br>AV: Adverb<br>AVP: Adverb particle(예, up, off, out)<br>AVQ: WH-adverb(예, when, why)<br>CJC: Conjunction<br>DT: Determiner(예, a/an, the, this, these)<br>NN: Noun<br>PNQ: WH-pronoun(예, who, whoever)<br>POS: The possessive (or genitive morpheme) 's or'<br>PRP: Preposition<br>VB: Verb<br>TO: Infinitive marker(예, to - infinitive)<br>DTQ: WH-determiner(예, whose, which)</td></tr>
</table>

〔그림 1〕 CLAWS Web Tagger를 활용한 품사 태깅

CLAWS Web Tagger는 무료 품사 태깅 프로그램으로서 두 종류(c5, c7)의 태그 세트(tag set) 중 하나를 선택적으로 적용할 수 있는데 일반적으로 많이 사용되는 c5 태그 세트도 62개의 품사로 태깅하기 때문에 복잡하다. 예를 들면, 동사를 VVI(infinitive of lexical verb), VVD(past tense form of lexical verb), VVG(-ing form of lexical verb) 등으로 세분하여 태깅한다. 이에 본 연구에서는 [그림 1]에 제시된 바와 같이 62개의 품사를 13개의 품사로 통합하여 분석에 적용하였다.

# IV. 연구 결과 및 논의

## 1. 한국인 학습자의 쓰기능력별 어휘 사용 양상 비교

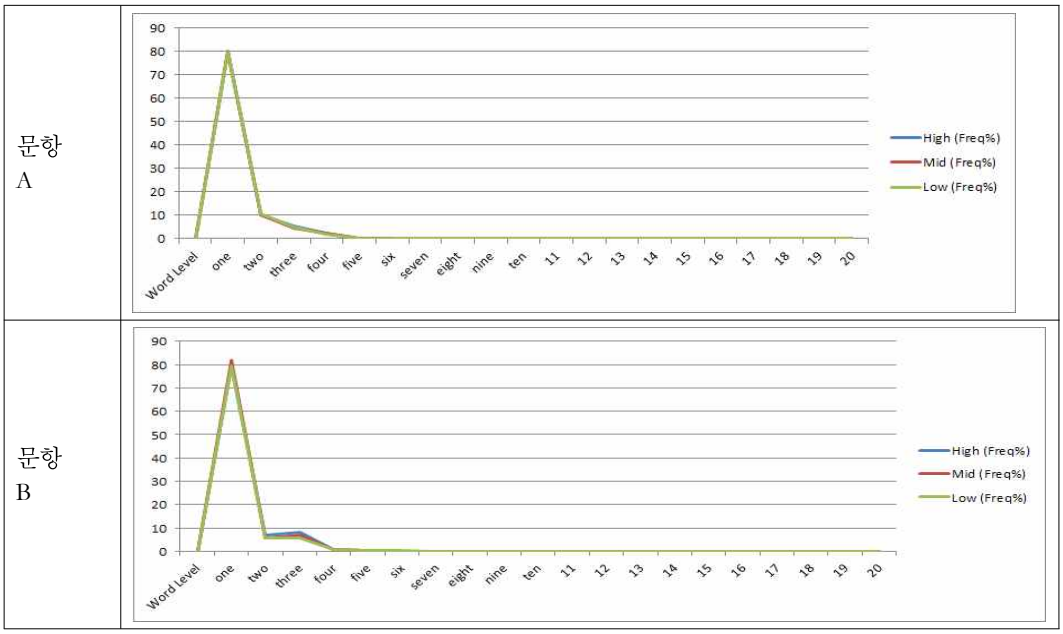
한국 대학생들의 영어 쓰기에 나타난 어휘 특성을 쓰기 수준별로 살펴보기 위해 BNC-COCA25 Range를 활용하여 쓰기 답안의 길이를 의미하는 총 단어 수(token), 어휘의 종류를 의미하는 유형 수(type), 텍스트 길이 대비 어휘의 다양성(Type-Token Ratio, TTR)을 분석하였다.

〔표 3〕 쓰기능력별 어휘 특성 비교

| 수준/문항 | | Token | Type | TTR |
|---|---|---|---|---|
| 상 | 문항 A | 20,102(M=402.04) | 2,468 | 0.12 |
| | 문항 B | 20,483(M=409.66) | 2,233 | 0.11 |
| 중 | 문항 A | 15,995(M=319.90) | 2,188 | 0.14 |
| | 문항 B | 16,464(M=329.28) | 1,883 | 0.11 |
| 하 | 문항 A | 10,093(M=201.86) | 1,625 | 0.16 |
| | 문항 B | 12,159(M=243.18) | 1,828 | 0.15 |
| 계 | 문항 A | 46,190(M=307.93) | 3,936 | 0.09 |
| | 문항 B | 49,106(M=327.37) | 3,689 | 0.08 |

[표 3]에서 볼 수 있듯이 문항 B에 대한 답안의 길이가 문항 A에 대한 답안에 비해 길었고 쓰기 점수가 높을수록 총 단어 수도 높게 나타났다. 이에 비해 어휘의 유형 수(type)는 문항 A의 경우 상위권일수록 많았지만 문항 B는 중위권(1,883)과 하위권(1,828)의 차이가 매우 근소하였다. 흥미롭게도 문항 B에 대한 답안의 총 단어 수가 A보다 더 많았음에도 불구하고 어휘의 유형 수는 하위권을 제외하면 A보다 낮은 것으로 나타났다. 이는 텍스트의 길이가 늘어나도 필수 단어들은 반복되어 유형 수는 줄기 때문이다(신동광 외, 2018).

어휘의 등급별 분석 결과는 [그림 2]에 제시된 바와 같이 상위 빈도 3,000 단어에 집중된 것을 알 수 있다. 어휘 사용의 절대 빈도는 상위권 학생들이 가장 높고 하위권 학생들이 가장 낮아서 세 집단의 차이가 명시적으로 나타났는데 등급별 어휘 사용 비율은 차이가 거의 없었다. 한편, 어휘의 다양성(TTR) 비율은 쓰기 수준이 높아질수록 점진적으로 낮아지는 패턴을 보였는데 이는 수준이 높아질수록 글이 길어지는데 기본 어휘는 반복적으로 출현하기 때문이다. 문항 A의 경우 학생들이 사용한 어휘의 90%가 상위 빈도 2,000 단어에 포함되었고 95%의 어휘가 상위 빈도 3,000 단어 이내에 포함된 것으로 나타났다. 문항 B에 대한 답안의 어휘 분포도 유사했는데 학생들이 사용한 어휘의 87%, 94%가 각각 상위 2,000 단어, 상위 3000 단어에 해당하는 것으로 나타났다. 이와 같은 결과는 한국 대학생들의 쓰기에 사용된 어휘의 93%가 상위 빈도 2,000 단어에 포함되고, 상위 빈도 3,000 단어 내에는 94%가 포함된다고 보고한 신동광 외(2018)의 연구 결과와 다르지 않다.




〔그림 2〕 쓰기 주제와 수준에 따른 어휘등급별 어휘 사용 분포 비율(%)

주제와 관련해서도 거의 유사한 분포를 나타냈으나 문항 B는 3rd 1000 등급에서 문항 A와 다른 분포를 보였다. 문항 B(상: 1684; 중: 1175; 하: 715)의 경우 문항 A(상: 1064; 중: 730; 하: 501)보다 총 단어 수(token)는 많았지만 어휘 유형 수(type)는 적어서(A상: 361; B상: 346; A중: 304; B중: 247; A하: 210; B하: 201) 결과적으로 문항 B의 TTR(상: 0.21; 중: 0.21; 하: 0.28)은 문항 A(상: 0.34; 중: 0.42; 하: 0.42)보다 낮게 나타났다. 문항 B에 대한 쓰기 답안의 경우 'media' (20회), 'method' (64회), 'communicate'의 파생·굴절 변이형(예: communicated, communicator, 813회) 등과 같은 3rd 1000 등급의 어휘들이 반복적으로 사용되어 불규칙한 분포를 보였다. 즉, 문항 B에서 3rd 1000 등급에 속하는 주제 관련 핵심어들이 반복적으로 사용되었기 때문에 전체 단어 수(token)는 많아도 유형 수(type)는 줄었고 결과적으로 어휘의 다양성(TTR)은 3rd 1000 등급에서 문항 A보다 낮아졌던 것으로 분석된다.

## 2. 한국인 학습자의 쓰기능력별 연어 사용 양상 비교

앞서 어휘 분석 결과에서 살펴보았듯이 문항 A와 B에 대한 답안에 나타난 어휘 분포 및 비율이 큰 차이가 없었기 때문에 연어 분석은 문항 A에 대한 150개 답안에 국한하여 이루어졌다.

### 1) 한국인 학습자의 쓰기능력별 연어 사용 양상 비교

본 연구에서는 한국 대학생들의 쓰기능력 수준별로 연어 특성을 살펴보기 위해 연어의 총 수(token), 유형 수(type), 총 수 대비 유형 수의 비율(TTR)을 [표 4]와 같이 분석하였다.

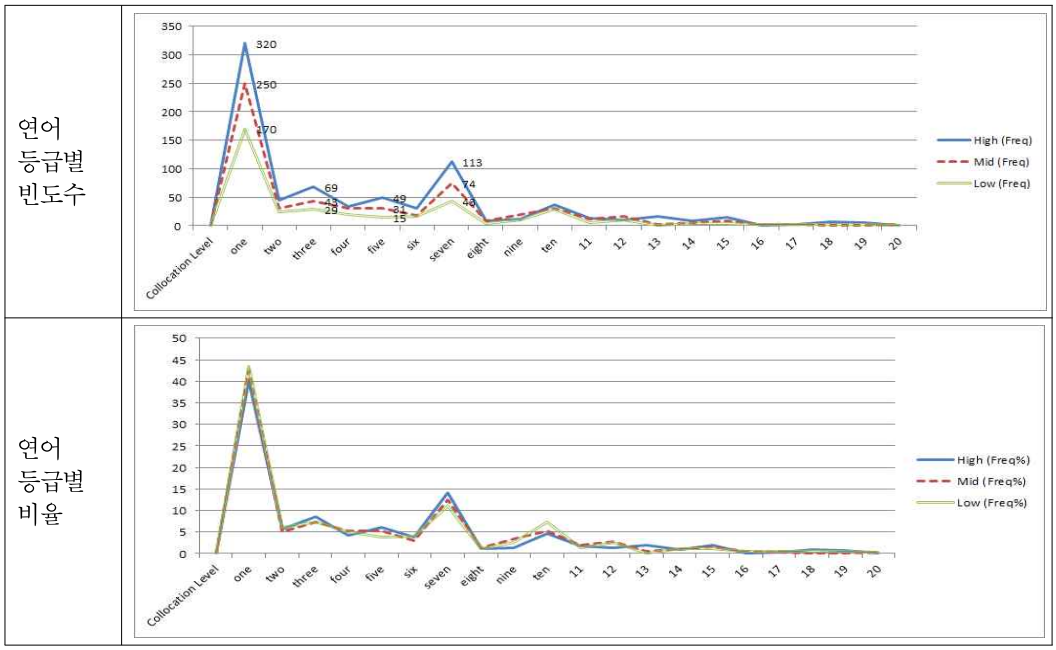
〔표 4〕 쓰기능력별 연어 특성 비교

| 수준 | Token | Type | TTR |
|---|---|---|---|
| 상 | 798(M=15.96) | 350 | 0.44 |
| 중 | 590(M=11.80) | 304 | 0.52 |
| 하 | 392(M=7.84) | 208 | 0.53 |
| 계 | 1,780(M=11.87) | 670 | 0.38 |

연어 사용 빈도는 상위권 학습자일수록 높았는데 상 수준(평균 15.96개)의 학생들은 중 수준(11.80개) 학생들과 하 수준(7.84개) 학생들보다 더 많은 수의 연어를 사용하는 것으로 나타났다. 연어의 유형수(type)도 수준이 높을수록 점진적으로 높게 나탔는데 상-하, 중-하의 차이가 상-중수준의 차이보다 좀 더 큰 것으로 나타났다. 한편 쓰기 답안의 길이 대비 연어 사용의 다양성을 측정한 TTR은 어휘에서와 같이 답안을 길게 작성한 상위 수준에서는 낮고 하위 수준에서는 높게 나타났다.

쓰기능력 수준에 따라 연어의 등급별 사용 빈도와 비율을 비교하면 [그림 3]과 같다. 구체적으로 살펴보면 연어 사용의 절대 빈도는 쓰기 답안의 길이가 상대적으로 긴 상위권에서 더 높았으나 연어 등급별 비율은 거의 유사하게 나타났다. 학생들의 글에서 1st 500 수준의 연어가 차지하는 비율은 전체 사용된 연어의 41.75%였고, 상위 빈도 2,000(상위 빈도 4개 등급) 수준의 연어가 커버하는 비율은 65.17%로 나타났다.

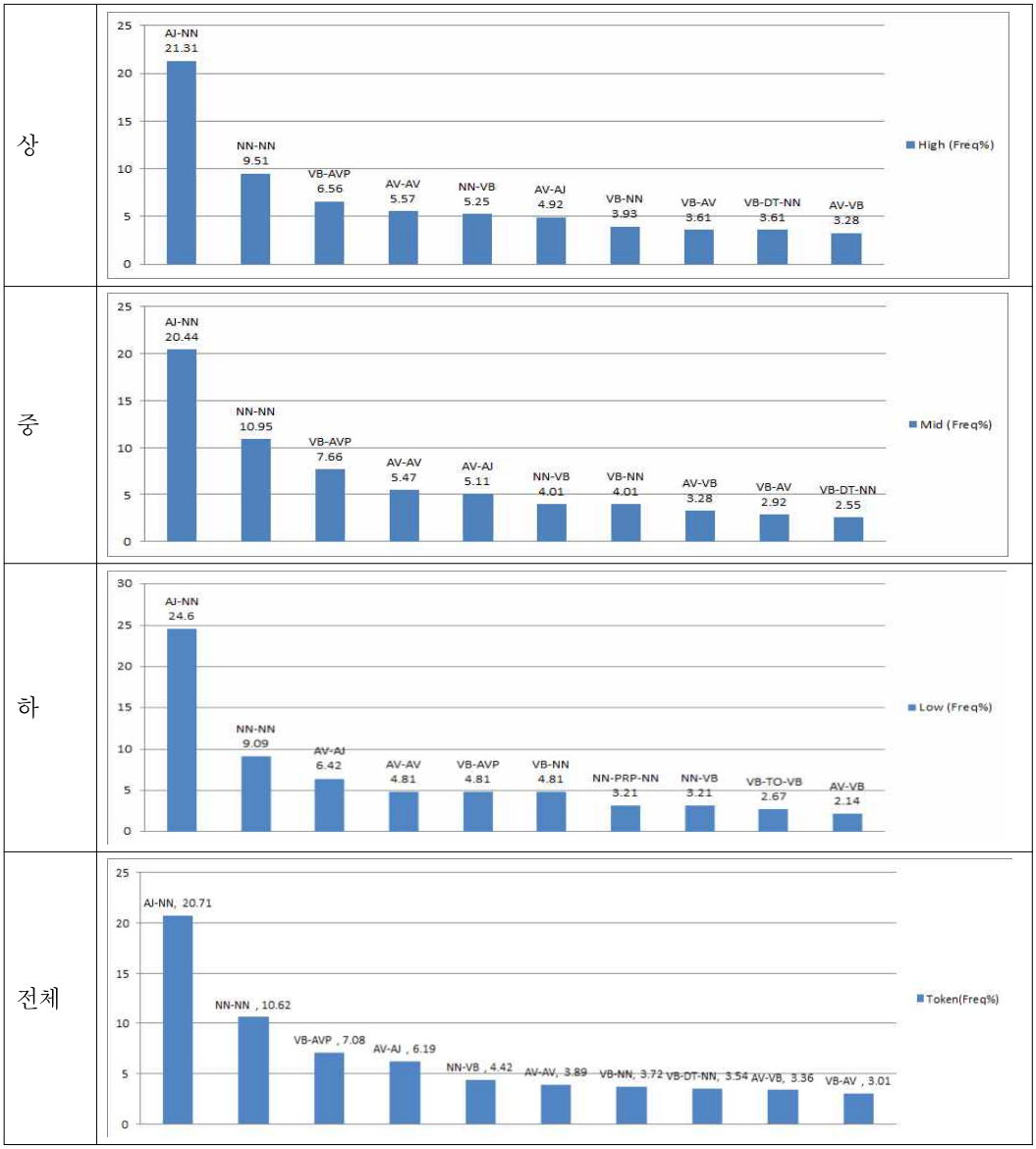
이와 같은 결과는 신동광 외(2018)가 보고한 한국인 대학생의 1st 500 수준의 연어 비율(46.68%), 상위 빈도 2,000(상위 빈도 4개 등급) 수준의 연어 비율(70.85%)보다 다소 낮다. 그러나 이는 신동광 외(2018)가 보고한 원어민의 1st 500 수준(38.14%), 상위 빈도 2,000(상위 빈도 4개 등급) 수준(64.24%)의 연어 비율에 근접하고 있어 본 연구의 참여자들이 상위 빈도 4개 등급 이상의 연어를 두루 두루 사용했음을 보여준다. 이를 통해 본 연구에 참여한 학생들이 다양한 수준의 연어를 사용하고 있으며 원어민의 영어 사용 양상과 유사한 것을 알 수 있다. 흥미롭게도 연어 사용 분포에서도 어휘에서와 같이 불규칙한 분포가 관찰되었다. 그 원인을 진단하기 위해 학생들이 사용한 연어들을 살펴보니 주제와 관련된 연어가 7등급 구간에 포함되었기 때문인 것으로 확인되었다. 예를 들어 '공인(公人)'을 뜻하는 'public figure'라는 연어 표현이 반복적으로(112회) 사용됨에 따라 불규칙한 분포가 나타났다.




〔그림 3〕 쓰기능력에 따른 연어 등급별 사용 분포

## 2) 한국인 학습자의 쓰기능력별 연어 품사조합 패턴 비교

전체 학생들의 연어 사용을 품사 조합 패턴 측면에서 분석한 결과 중복된 것을 제외하면 총 95개의 품사 조합 패턴을 사용한 것으로 나타났다. 이를 수준별로 나누어 분석한 결과 상위권 학생들과 중위권 학생들은 각각 61개, 65개의 연어 품사조합 패턴을 사용한데 반해 하위권 학생들은 55개 사용한 것으로 나타났다. 상위 수준과 중위 수준의 학생들이 사용한 품사 패턴 수의 차이는 적었지만 품사패턴의 사용 빈도는 상위 수준(305회)이 중위 수준(274회), 하위 수준(187회)보다 높은 것으로 나타났다. [그림 4]는 상·중·하 수준과 전체의 연어 품사 조합 패턴을 비교한 것이다.




〔그림 4〕 쓰기능력별 연어 품사조합 패턴 (상위 10개)

연어의 전체 패턴을 살펴보면 쓰기능력에 상관없이 AJ-NN 품사 조합 패턴(20% 이상)이 가장 많이 사용되었고 NN-NN(10% 내외) 패턴이 그 뒤를 이었음을 알 수 있다. 상위 빈도 10개의 연어 품사 조합 패턴을 분석한 결과, 상위권과 중위권 학생들의 품사 패턴은 순위에서는 차이가 있었지만 10개 패턴의 종류는 차이가 없었다. 또한 품사 패턴별 사용 비율에 있어서도 상위 학생과 중위 학생 간 차이가 근소한 것으로 나타났다. 이에 반해 하위권의 경우 상위권과 하위권의 10개 패턴에 포함된 VB-AV(예: go there), VB-DT-NN(예: breaking the law)등의 품사 조합 대신 NN-PRP-NN(예: kind of thing),

VB-TO-VB(예, want to know) 패턴을 사용하고 모든 품사조합 패턴에서 상위, 중위 학생들보다 연어 사용 비율이 낮은 것으로 나타났다.

본 연구에서 관찰된 품사 패턴을 선행연구들과 비교해보면 Lewis(2000)가 제안한 여섯 가지 연어 패턴 중 NN-VB-PRP(예: fog closed in) 조합을 제외한 다섯 가지 패턴이 본 연구의 연어 품사조합 패턴 상위 10개에 포함되었다. 이와 같은 양상은 McCarthy와 O'Dell(2005)이 제시한 품사 패턴과의 비교에서도 관찰되었다. 구체적으로 VB-PRP-NN(예: filled with horror) 조합을 제외한 다섯 개의 패턴이 본 연구의 연어 품사조합 상위 10개에 포함되었다. McCarthy와 O'Dell(2005), Lewis(2000)가 제시한 전체 주요 연어 패턴을 통합하여 본 연구의 품사 패턴 상위 10개와 매칭해보니 AJ-NN(예: bright color), NN-NN(예: radio station), AV-AJ(예: extremely inconvenient), NN-VB(예: economy boomed), VB-AV(예: smiled proudly), VB-DT-NN(예: submit a report), AV-VB(예: happily married) 등의 일곱 가지 패턴이 일치하는 것으로 나타났다.

## V. 결론 및 제언

본 연구는 서울 소재 대학에서 신입생을 대상으로 실시한 최근 5년간의 진단평가 자료로 구축된 영어학습자 코퍼스(HELC) 데이터 중 2018년 평가 자료에 기초하고 있다. 온라인으로 진행된 2018년 진단평가에서 두 개의 쓰기 문항(문항 A, 문항 B) 중 한 문항이 응시자에게 임의 배정되었는데, 문항별로 상, 중, 하 세 개의 수준에 해당하는 50명의 답안을 무작위로 추출하여 각 150개씩 총 300개의 답안 데이터를 구축하고 이를 대상으로 어휘 수준, 분포, 다양성 등을 분석하였다. 쓰기 수준별로 어휘 사용 양상을 비교한 결과, 상위 수준의 경우 어휘의 절대빈도가 높은데 반해 하위 수준의 경우 어휘의 절대빈도가 낮은 것을 알 수 있었다. 그러나 어휘등급별 분포에서는 상, 중, 하 수준에 관계없이 유사한 분포를 보였다. 또한, 쓰기 주제에 관계없이 전체 어휘의 약 90% 정도가 상위 2000단어에 포함되었고, 약 95% 정도가 상위 3000단어 수준에 포함되는 것으로 나타났다.

한편, 연어 사용 분석은 문항 A에 대한 쓰기 답안 150개에 초점을 두고 이루어졌는데, 어휘 분석 결과와 유사한 패턴이 관찰되었다. 즉, 연어의 절대 사용빈도가 상위권에서 높게 나타났으나 쓰기 수준별 연어 사용 비율은 큰 차이가 없었다. 연어 품사조합 패턴과 관련해서도 상위 10개 패턴의 순위에서 차이가 관찰되었으나 근소했다. 상 수준과 중 수준은 10개 패턴의 종류가 일치하였고 하 수준과는 9개가 일치하였다. 흥미롭게도 AJ-NN(20% 이상), NN-NN(10% 내외), VP-AVP(7%) 등의 분포를 보여 특정 품사 패턴에 편중되는 것으로 나타났다.

본 연구는 한국인 대학생들의 어휘와 연어 사용 양상을 쓰기 수준별로 비교했다는 점에서 의의가 있다. 분석 결과 어휘와 연어 사용의 절대 빈도가 학습자의 쓰기 수준이 높아질수록 높아졌지만 어휘와 연어 사용 비율 면에서는 유사한 분포를 나타내는 것을 확인하였다. 주목할 것은 90%의 어휘가 상위 2000단어 등급에 포함되어 있다는 점, 어휘와 비교했을 때 연어의 총 수와 유형 수가 낮은 점, 연어

의 품사 패턴이 특정 품사 조합에 편중된다는 점이다.

이와 같은 결과는 교육적 함의가 있는데 우선 어휘와 관련해서 상위 2000단어, 3000 단어에 편중되어 있는 문제를 해결하기 위한 방법을 연구해야 한다. 현장 교사들은 학생들이 단순히 듣기, 읽기 등의 이해 기능에서 어휘를 인지하는 데 그치지 않고 언어 산출(language production)의 과정을 통해 어휘를 직접적으로 사용할 수 있는 기회를 제공해야 할 것이다. 이를 위해 읽기 연계 글쓰기 과업을 설계하여 학생들에게 쓰기 과업을 수행하게 함으로써 이해어휘가 표현어휘로 전이될 수 있도록 해야 할 것이다. 예를 들어, 학생들에게 영어 글감을 제시하여 읽게 한 후 이를 패러프레이즈하거나 요약하게 하면서 원문의 단어와 의미가 유사하거나 동일한 어휘로 대체하는 활동을 진행할 수 있다. 또한 사전 쓰기 활동으로 주제 관련 어휘 불러오기 활동을 실시하여 표현어휘에 대한 의식을 강화할 수 있다. 무엇보다도 등급별로 다양한 어휘에 노출될 수 있는 기회를 마련하는 것이 중요한데 이를 위해 교사는 등급별로 어휘를 무작위로 선정하여 이를 가지고 스토리를 써보게 하는 등의 활동을 구안, 적용할 수 있다.

한국 학생들에게 어휘의 균형 있는 발달도 중요한 과제이지만 연어 습득은 더욱 시급한 과제이다. 비원어민 학습자들이 연어 사용을 어려워하고 제한된 패턴의 연어만 사용한다는 것은 널리 알려진 사실이다(Laufer & Waldman, 2011; Nesselhauf, 2005; Siyanova-Chanturia, 2015; Vedder & Benigno, 2016). Siyanova-Chanturia는 비원어민의 영어는 어색하고 부자연스러운데 반해 원어민 영어는 실제성이 높고 자연스럽다고 소개하면서 그 차이를 유발하는 요인이 연어라고 설명하였다. 많은 연구들을 통해 연어의 중요성이 강조되었는데 이는 연어지식이 어휘 발달의 질적 지표로서 어휘 지식을 구성하기 때문이다(Bahns & Eldaw, 1993).

따라서 영어과 교육과정 개발자들은 어휘 지식을 구성하는 요소로서 연어의 중요성을 인지하고 연어를 교육과정의 내용체계에 반영해야 할 것이다. 어휘에 대해 교육과정 기본어휘라는 어휘목록이 있듯이 연어의 난이도를 분석하여 영어과 교육과정 내에서 연어목록을 제시해야 할 필요가 있다. 교육과정의 원리 및 지침에 따라 교재 개발자들은 영어 교과서 내용 중 연어의 비율을 늘려야 하며 다양한 패턴의 연어 표현을 수록해야 할 것이다. 또한 현장교사들은 매우 제한된 수의 '숙어' 표현과는 다른 연어를 효과적으로 가르칠 수 있는 교수 방법 및 과업을 구안해야 할 것이다. 뿐만 아니라 학생들에게 연어를 검색할 수 있는 방법을 소개하는 것을 포함하여 연어를 왜 배워야 하는지에 대한 동기를 부여하기 위한 노력을 해야 한다. 예를 들어 비원어민 영어와 원어민 영어의 차이를 유발하는 요인 중 하나가 연어 사용임을 실증적으로 보여주어 학습자가 연어 습득의 필요성을 공감할 수 있도록 하는 노력이 필요하다(Siyanova-Chanturia, 2015).

## 참고 문헌

신동광. (2018). 영어 입력 모드의 차이에 따른 요약쓰기의 어휘적 특성 비교. *외국학연구, 46*, 39-66.

신동광, 배주경, 송민영. (2014). 영어 쓰기 답안에 나타난 한국 고등학생들의 단어 조합 오류 유형 분석. *Studies in English Education, 19*(2), 261-282.

신동광, 전유아, 이신웅, 박명수. (2018). 한국인 학습자와 영어 원어민의 구어 및 문어 코퍼스에 나타난 개별 어휘 및 다어휘 표현 비교·분석. *영어교과교육, 17*(2), 93-112.

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics, 22*(2), 173 - 195.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System, 21*(1), 101 - 114.

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System, 69*, 65-78.

Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language, 22*(1), 181-200.

Cobb, T. (2012). Compleat lexical tutor version 6.2 [Computer Software]. Available from http://www.lextutor.ca

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Coxhead, A. (2016). Reflecting on Coxhead (2000), "A new academic word list." *TESOL Quarterly, 50*(1), 181-185.

Cross, J., & Papp, S. (2008). Creativity in the use of verb + noun combinations by Chinese learners of English. In G. Gilquin, S. Papp & M. B. D´ıez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 57 - 81). Amsterdam, Netherlands: Rodopi.

Dang, T., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes, 33*, 66-76.

Douglas, S. R. (2013). The lexical breadth of undergraduate novice level writing competency. *The Canadian Journal of Applied Linguistics, 16*(1), 152-170.

Durant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching, 47*, 157-177.

Ehrman, B., & Warren, B. (2000). The idiom principle and the open-choice principle. *Text & Talk, 20*(1), 29-62.

Ghadessy, M. (1989). The use of vocabulary and collocations in the writing of primary school students in Singapore. In I. S. P. Nation & R. Carter (Eds.), *Vocabulary acquisition* (pp. 110-117). AILA Review, No. 6.

Gregori-Signes, C., & Clavel-Arroita, B. (2015). Analysing lexical density and lexical

diversity in university students' written discourse. *Procedia-Social and Behavioral Sciences, 198*, 546-556.

Gyllstad, H. (2009). Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLMATCH. In A. Barfield & H. Gyllstad (Eds.), *Researching second language collocation knowledge* (pp. 153-170). New York: Palgrave Macmillan.

Imura, M. (2002). *A study on corpus-based teaching of colloquial English: development of application of cinema transcripts database* (Unpublished doctoral dissertation). Osaka University, Osaka, Japan.

Kao, S. M., & Wang, W. C. (2014). Lexical and organizational features in novice and experienced ELF presentations. *Journal of English as a Lingua Franca, 3*(1), 49-79.

Kashiha, H. & Chan, S. H. (2015). A little bit about: Differences in native and non-native speakers' use of formulaic language. *Australian Journal of Linguistics, 35*(4), 297-310.

Kiliç, M. (2019). Vocabulary knowledge as a predictor of performance in writing and speaking: A case of Turkish EFL learners. *PASAA, 57*, 133-164.

Kim, S. Y., & Le, T. H. (2018). A corpus analysis of collocational behaviors of near-synonymous adjectives. *Multimedia-Assisted Language Learning, 21*(4), 181-210

Kim, S. Y., & Ryu, Y.-S. (2009). Korean college students' vocabulary profiles as predictors of English reading and writing proficiency. *Multimedia-Assisted Language Learning, 12*(3), 93-115.

Kim, S. Y., & Ryu, Y.-S. (2011). Korean EFL Learners' vocabulary use in reading-based writing. *영어교육, 66*(1), 91-109.

Kuo, C. (2009). An analysis of the use of collocation by intermediate EFL college students in Taiwan. *ARECLS, 6,* 141-155.

Laufer, B., & Hulstijin, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*(1), 1-26.

Laufer, B., & Nation, I. S. P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 6*(1), 33 - 51.

Laufer, B., & Nation, I. S. P. (2001). Passive vocabulary size and speed of meaning recognition: Are they related? In S. Foster-Cohen & A. Nizegorodcew (Eds.), *EUROSLA yearbook 1* (pp. 7 - 28). Amsterdam, Netherlands: Benjamins.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647-672.

Lee, Y., Chon, Y. V., & Shin, D. (2012). Vocabulary size of Korean EFL university learners: Using an item response theory model. *English Language & Literature Teaching, 18*(1), 171-195.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British national corpus*. London: Longman.

Lewis, M. (1993). *The lexical approach*. Hove, UK: Language Teaching Publications.

Lewis, M. (Ed.). (2000). *Teaching collocation: Further developments in the lexical approach*. Hove, UK: Language Teaching Publications.

Li, J., & Schmitt, N. (2010). The development of collocation use in academic texts by advanced L2 learners: A multiple case study approach. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 2-46). London, UK: Continuum.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320.

McCarthy, M., & O'Dell, F. (2005). *English collocations in use* (5th ed.). Cambridge: Cambridge University Press.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 57-78). EuroSLA Monograph 2. The European Second Language Association.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P. (2012). *The BNC/COCA word family lists* [Data File]. Available from http://www.victoria.ac.nz/lals/about/staff/publications/BNC_COCA_25000.zip

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59-82.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam, Netherlands: John

Benjamins.

Nurmukhamedov, U. (2017). The contribution of collocation tools to collocation correction in second language writing. *International Journal of Lexicography, 30*(4), 454 - 482.

Paquot, M. (2007). Towards a productively-oriented academic word list. In J. Walinski, K. Kredens & S. Gozdz-Roszkowski (Eds.), *Practical applications in language and computers 2005* (pp. 127 - 140). Frankfurt am Main, Germany: Peter Lang.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 174-199). Cambridge: Cambridge University Press.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 355-371.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*, 329 - 363.

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching, 50*(2), 212-226.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26-43.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88.

Shimamoto, T. (2000). An analysis of receptive vocabulary knowledge: Depth versus breadth. *JABAET Journal, 4*, 69-80.

Shin, D. (2015). The influence of vocabulary regulations in the 2009 national curriculum of English on English textbooks. *The Journal of Foreign Studies, 34*, 47-76.

Shin, D., Chon, Y. V., Lee, S., & Park, M. (2018). COCA_MWU20 ColloGram [Computer Software].Available from http://cfile281.uf.daum.net/attach/99EBA0495A80F110304D74

Shitu, F. M. (2015). Collocation errors in English as a second language (ESL) essay writing. *International Journal of Social, Behavioral, Educational, Economic, Business, and Industrial Engineering, 9*(9), 3270-3277.

Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System, 53*, 148-160.

Siyanova-Chanturia, A., & Schmitt, N. (2008). L2 learner production and processing of

collocation: A multi-study perspective. *The Canadian Modern Language Review, 64*, 429-458.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal, 36*(2), 139-152

Tsai, K.-J. (2015). Profiling the collocation use in ELT textbooks and learner writing. *Language Teaching Research, 19*(6), 723-740.

University Centre for Computer Corpus Research on Language (UCREL). (2014). Free CLAWS Web Tagger [Computer Software]. Available from http://ucrel-api.lancaster.ac.uk/claws/free.html

Vedder, L., & Benigno, V. (2016). Lexical richness and collocational competence in second language writing. *International Review of Applied Linguistics in Language Teaching, 54*(1), 23-42.

West, M. (1953). *A general service list of English words*. London, UK: Longman.

Yüksel, İ. (2015). The key to second language writing performance: The relationship between lexical competence and writing. *Sino-US English Teaching, 12*(8), 539-555.

Applicable levels: secondary, tertiary education

Authors: Kim, Sung Yeon (Hanyang University, 1st author); sungkim@hanyang.ac.kr
Shin, Dongkwang (Gwangju National University of Education, corresponding author); sdhera@gmail.com
Kim, Kyung-Sook (Hanyang University, co-author); cindytesol@hanyang.ac.kr

Received: April 30, 2020
Reviewed: May 20, 2020
Accepted: June 15, 2020