http://dx.doi.org/10.18649/jkees.2018.17.2.93

한국인 학습자와 영어 원어민의 구어 및 문어 코퍼스에 나타난 개별 어휘 및 다어휘 표현 비교·분석*

신동광 · **전유아** · **이신웅**** · **박명수**
광주대학교 · 한양대학교 · 한양대학교 · 상명대학교

Shin, Dongkwang, Chon, Yuah, Lee, Shinwoong, & Park, Myongsu (2018). A comparison of single word and multi-word unit profiles in spoken and written corpora of Korean learners and English native speakers. ***Journal of the Korea English Education Society, 17*****(2), 93-112.**

The present study aimed to compare English non-native speaker corpora vs. English native speaker corpora for the analysis of spoken and written English. Programs BNC-COCA 25000 RANGE and COCA_MWU20 ColloGram were used to analyze frequency and lexical variety of single words and multi-word units (MWUs) in each of the four corpora: Korean EFL learners' spoken corpus and written corpus, BNC Spoken Sampler and BNC Written Sampler. The analysis of the four corpora demonstrated: First, the Korean learners' use of single word items and MWUs in the spoken and written English did not show a noticeable difference compared to the native speakers' use of the same lexical items. Second, more than 90% of the single words consisted of items from the first 5,000 words whereas 70% of the MWUs were made up of those from the first 2,500 MWUs. Third, similar to previous studies, there was the repeated use of single words and MWUs in spoken English while a wider range of expressions were used in written English. Fourth, the results of the analysis indicated that the Korean learners were exposed to different opportunities for English vocabulary learning compared to native speakers, which may lead to Korean learners' inefficient learning of vocabulary. Implications are further discussed for improved vocabulary learning.

[BNC-COCA 25000/COCA_MWU20/Korean learner corpus/ BNC-COCA 어휘 목록/COCA 다어휘 표현 목록/학습자 코퍼스]

---

* 이 논문은 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2015S1A5A2A03048887).

** 제1저자
*** 교신저자

I. 서론

코퍼스라는 언어 빅데이터가 영어교육 및 관련 연구에 활발히 적용되면서 개별 어휘 뿐 아니라 두 단어 이상으로 구성된 다어휘 표현(Multi-Word Unit, MWU) 또한 그 중요성이 주목 받게 되었다(Bahns & Eldaw, 1993; Lewis, 2000). 단순한 반복적인 패턴에 초점을 둔 'n-gram'과는 달리 다어휘 표현은 빈도수(frequency), 사용범위(range), 문법성(grammatical well-formedness), 의미의 투명성(semantic transparency) 등 다양한 기준을 적용하여 선정한 유용한 표현을 의미하며 이를 연구하는 학자들에 따라 'collocation,' 'idiom,' 'lexical bundle,' 'formulaic sequence' 등 다양한 용어로 지칭되었다(Nation, 2016).

이제 개별 어휘뿐만 아니라 다어휘 표현은 제 2언어 또는 외국어 학습에서 매우 중요한 학습 요인이며 언어 사용의 정확성과 유창성에서도 중요한 역할을 한다는 연구가 다수 발표되었다(예, Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006). 이와 관련하여 많은 해외 연구들은 대부분 영어 원어민 코퍼스 분석을 통해 소규모의 다어휘 표현 목록을 개발하거나 특성을 분석하는데 초점을 두고 수행되었다(Biber, Conrad & Cortes, 2004; Liu, 2003). 반면 국내에서는 대부분의 연구가 다어휘 표현의 학습지도 방법 및 평가에 그 연구가 한정되었다(김혜선, 윤현숙, 2010; 유경미, 김낙복, 2014; 이재근, 김은주, 2008). 이는 한국인 학습자를 대상으로 한 코퍼스 구축이 상당한 시간이 걸리고 특히 구어 코퍼스 구축에는 많은 인력과 비용이 소요된다는 점에서도 그 원인을 찾을 수 있다. 최근 국내 여러 대학에서는 한국인 학습자를 대상으로 한 코퍼스 구축이 활발하다(예, 연세대학교 YELC). 그럼에도 불구하고 대부분의 학습자 코퍼스가 특정 주제에 한정된 문어 자료 기반으로 구축되었기 때문에 한국인 학습자의 개별 어휘나 다어휘 표현의 사용을 구어와 문어로 나누어 비교·분석하거나 개별 어휘와 다어휘 표현 사용 간의 관계성 분석에 관한 연구는 여전히 찾아보기 힘들다.

이에 따라 본 연구에서는 대표성이 확보된 등급화된 개별 어휘 목록과 다어휘 표현 목록을 탑재한 측정 도구를 활용하여 영어 원어민 코퍼스와 한국인 학습자 코퍼스를 구어와 문어 자료로 구분하여 분석하고 각 코퍼스에서 나타나는 개별 어휘 및 다어휘 표현 양상을 다각도로 비교·분석하고자 하였다. 다음은 본 연구에서 설정한 연구 질문이다.

첫째, 한국인 학습자/영어 원어민 그리고 구어/문어의 특성이 반영된 코퍼스들에서 개별 어휘 및 다어휘 표현의 사용 빈도에는 어떠한 차이가 있는가?

둘째, 한국인 학습자/영어 원어민 그리고 구어/문어 특성이 반영된 코퍼스들에서 개별 어휘 및 다어휘 표현 사용의 다양성에는 어떠한 차이가 있는가?

셋째, 위의 연구 질문에서 제시된 변인이 개별 어휘 및 다어휘 표현 사용의 차이에 영향을 주었다면 그 원인과 교육적 시사점은 무엇인가?

II. 이론적 배경

1. 구어와 문어 코퍼스에서의 개별 어휘/다어휘 표현 사용 양상 분석

Halliday(1985)는 구어와 문어는 같은 표현에 대해 단순히 표현 방식이 다른 것이 아니라 그 표현 자체가 다른 것이라고 말할 정도로 두 가지 모드(mode)의 차이점을 강조하였다. McEnery와 Hardie(2012)는 전통적으로 언중(言衆)이 사용하는 표현이나 문법은 문어 사용에 기반하고 있었으나 보다 많은 언어학자들이 구어만의 독특한 표현이나 문법 체계에 대해 관심을 가지고 이를 밝히기 위한 연구를 수행하고 있다고 주장하였다.

Leech, Rayson과 Wilson(2001)의 연구에서는 영국의 대표적인 코퍼스인 British National Corpus(BNC)를 바탕으로 구어와 문어 자료에 나타난 개별 어휘의 빈도, 사용범위, 분포빈도수, 품사분포 등의 정보를 분석하여 그 차이를 비교하였다. Lee(2001) 또한 BNC Sampler Corpus에 포함된 다양한 영역과 장르의 개별 어휘 분포를 분석하였다. Lee의 연구에 따르면 구어 자료에서는 매우 제한된 수의 개별 어휘만이 사용되었고 이들 개별 어휘의 대부분은 기초 핵심 개별 어휘 목록(상위 빈도 2000단어)에 포함되었다. 또한 개별 어휘의 다양성을 나타내는 TTR(Type Token Ratio)의 경우 구어에서 .02, 문어에서는 .05로 문어 자료에서 더 다양한 개별 어휘가 사용되었으며 전체 개별 어휘 사용 면에서는 기초 핵심 개별 어휘 목록이 차지하는 비율이 구어에서는 90.52%, 문어에서는 74.90%로 나타났다. Imura(2002)의 연구에서는 구어에서 나타나는 개별 어휘에 주목하였는데 이를 위해 그는 영화 대본에 나타난 개별 어휘 사용의 특성을 분석하였다. 그 결과 구어에서는 특정 다어휘 표현이나 대명사의 사용이 일반 문어 자료에 비해 압도적으로 많았고 줄임말(예, gonna, dunno)이나 간투어(예, erm, well), 담화표지(예, so, thus)도 자주 사용되는 것으로 나타났다. 이를 바탕으로 그는 구어 모드에서 구사되는 표현은 문어 표현과 구분하여 지도할 필요가 있다고 강조하였다.

구어와 문어 코퍼스에 나타난 다어휘 표현의 비교는 개별 어휘 관련 연구보다는 그 수가 매우 제한적이다. 대표적인 연구로 Erman과 Warren(2000)은 구어와 문어 코퍼스에서 사용되는 다어휘 표현(prefab)의 사용 특성을 비교하였다. Erman과 Warren의 연구에서는 각 5,000단어 정도로 구성된 구어와 문어 자료를 분석하였고 그 결과, 전체 사용된 총 어휘 수에서 다어휘 표현이 차지하는 비중은 구어에서 약 59%, 문어에서는 약 52%로 나타나 구어에서 다어휘 표현이 더 빈번하게 사용된다는 것을 알 수 있었다. 또한 다어휘 표현의 평균 길이는 구어에서 2.61단어, 문어에서 2.80단어로 문어에서 더 긴 표현을 사용한다는 것도 확인할 수 있었다. 국내에서는 Shin(2007)이 영어 원어민의 구어와 문어 코퍼스에 나타난 다어휘 표현(collocation)을 비교·분석하였다. 그의 연구는 구어와 문어 각 1,000만 단어의 코퍼스를 분석하여 비교적 대규모 코퍼스를 활용한 연구

였다. 그의 연구에서는 상위 빈도의 50개 구어와 문어 다어휘 표현을 비교하였고 그 결과, 두 목록에 공통적으로 포함된 표현은 불과 15개에 불과하였다. 또한 두 코퍼스에 사용된 전체 다어휘 표현의 수는 유사했으나 문어 코퍼스와 비교하여 구어 코퍼스에서는 소수의 표현이 보다 집중적으로 반복되어 사용되는 것을 확인할 수 있었다. 즉 구어에서 다어휘 표현은 문어에서 보다 중요한 역할을 한다는 것을 알 수 있다. 기존 연구(예, Bresnan, 1999)에서도 말하기와 같은 구어 모드는 실시간(real time)으로 이루어지며 이에 따라 대부분 비동시적(delayed time) 활동으로 이루어지는 쓰기 보다는 작업 기억(working memory)의 부담이 커지는 만큼 사전에 이미 조합된 형태(ready-made sequence)로 인지적 부담을 덜어주는 다어휘 표현에 대한 의존도가 커진다고 주장하였다.

2. 비원어민 학습자와 영어 원어민의 개별 어휘/다어휘 사용 양상 분석

언어 노출 환경 그리고 모국어 간섭 등의 이유로 비원어민의 제 2언어 또는 외국어 구사는 원어민의 모국어 구사와 분명한 차이를 보인다. 1920-30년대 일본에서 활동한 Palmer(1925)는 일본인 학생들에 대해 다음과 회고한 바 있다: “일본인 학생들은 인내심을 가지고 참 열심히 단어나 표현을 암기한다. 하지만 그 개별 어휘나 표현들 가운데는 영어 원어민에게도 생소한 쓸데없는 표현들이 많았다. 이것이 내가 일본에서 경험한 가장 슬픈 일중 하나였다”(p. 190). 이는 언어 학습의 목적과 언어 노출 환경의 차이가 비원어민 화자의 언어 학습에 어떠한 영향을 미치는지를 보여주는 사례라고 할 수 있다.

Shirato와 Stepleton(2007)은 영어교육의 시사점을 도출하기 위해 일본인 학습자 코퍼스를 기반으로 구어와 문어 측면에서 영어 원어민 코퍼스에서 나타난 개별 어휘 및 다어휘 표현 사용과 어떠한 차이가 있는지를 분석하였다. 특히 이들은 일본인 학습자들이 영어 원어민 화자와 비교하여 어떤 개별 어휘 또는 다어휘 표현을 더 빈번히 사용하는지 또는 덜 사용하는지에 초점을 두고 살펴보았다. 그 결과, 일본인 학습자들은 많은 부분에서 차이를 보였는데 그 중에서도 표현의 모호성을 구사하기 위해 인과적 명확성을 나타내는 담화표지나 상호작용 시 우회적이거나 정중하고 부드러운 표현을 나타내는 어휘들의 사용 빈도는 낮은 반면 일부 빈도가 높은 단어나 조동사를 과도하게 사용하는 경향을 나타냈다. 또한 개별 어휘 사용에 있어 어휘의 기본 의미에 사용이 집중되고 다양한 의미나 다어휘 표현과 같은 다양한 활용 면에서는 미숙함을 드러냈다. 이러한 특징은 국내 연구에서도 나타난다. 신동광(2015)의 연구에 따르면 교사발화의 분석에서 원어민 교사들의 발화와는 달리 한국인 교사들은 정중한 요청의 표현(예, would you~, could you~) 대신 명령법이나 ‘please’를 추가하는 정도의 표현을 주로 구사한다고 지적하기도 하였다. Shirato와 Stepleton는 이러한 비원어민의 특성은 주로 문어 표현 위주로 교육을 받아 구어적 특성에 대해 익숙하지 않고 구어에서 자주 사용되는 다어휘 표현 보다는 개별 단어의 학습에 집중하

는 경향이 있기 때문이라고 주장하였다.

Kashiha와 Chan(2015)은 교실 내 토론 활동에서 말레이시아인 학습자와 영어 원어민 화자가 구사하는 다어휘 표현(lexical bundle)의 유형을 비교하였다. 이를 위해 4개 단어로 구성된 다어휘 표현(4 word sequence)을 토론 전사 자료(transcript)에서 추출하였고 이를 담화 기능별로 비교하였다. 그 결과 원어민 화자들은 말레이시아인 학습자들 보다 빈번하게 다어휘 표현을 구사하였고 담화를 시작하거나 부연설명에 필요한 표현들(예, want to talk about, on the other hand)에서 더 높은 빈도수를 보였다. 반면 말레이시아 학습자들은 가능성이나 태도 등을 나타내는 표현(예, is likely to be, do you want to)에서 더 높은 빈도수를 보였다. 예를 들어, 말레시아인 학습자들은 'agree with'와 같이 소수의 특정 다어휘 표현을 과도하게 사용하는 경향을 보였지만 원어민 화자가 자주 구사하는 담화표지, 비교구조의 표현, 그리고 다양한 기능을 포함하는 표현 등은 사용빈도가 매우 낮게 나타났다. 즉 비원어민 화자의 발화는 단순한 표현을 반복적으로 사용하는 경향을 보였다. Kashiha와 Chan은 이러한 언어 구사의 한계가 다어휘 표현에 대한 낮은 노출 빈도에서 기인한다고 보았다.

Ishikawa(2015)는 기존의 연구와는 다소 차이가 있는 연구결과를 보여주었다. Ishikawa의 연구에서는 주제가 통제된 구어와 문어 자료를 바탕으로 일본인 학습자와 영어 원어민 화자의 개별 어휘 사용을 비교하였다. 그 결과 원어민의 개별 어휘 구사에서는 구어와 문어 특성 면에서 유의한 차이를 보이지 않았지만 일본인 학습자의 경우에는 상대적으로 원어민 보다는 큰 차이를 보였다. 그럼에도 불구하고 Ishikawa는 이러한 차이는 일반적으로 예상하는 차이보다 작았으며 이러한 결과를 바탕으로 특히 초급이나 중급 수준의 비원어민에게는 구어와 문어의 표현을 구분하기 보다는 통합적인 표현에 집중하는 방식의 교수·학습이 적절하다고 제안하였다.

3. 개별 어휘와 다어휘 표현 사용의 관계성 분석

개별 어휘와는 달리 다어휘 표현에 대해서는 원어민 화자라고 할지라도 인식하지 못하는 경우가 많으며 실제 비원어민은 물론 원어민들도 듣기나 읽기에서는 다어휘 표현을 하나의 표현 단위라고 인식할 필요성을 느끼지 못하는 경우가 많다. 하지만 말하기와 쓰기와 같은 표현 능력에서는 이러한 다어휘 표현 지식이 언어 구사 능력에 큰 영향을 미친다. 이러한 이유로 개별 어휘에 대한 지식의 수준과 다어휘 표현의 지식수준이 항상 일치한다고 보기는 어렵다(Bahns & Eldaw, 1993). 보통 이를 두 가지 지식의 관계성을 분석하기 위해서는 개별 어휘 평가지와 다어휘 표현 평가지를 활용하여 그 결과의 상관성을 분석하는 것이 일반적이다.

Mutlu와 Kaşlıoğlu(2016)는 터키인 학습자들을 대상으로 Nation과 Beglar(2007)가 개발한 절대적인 개별 어휘 지식 평가지인 'Vocabulary Size Test(VST)'을 활

용하여 개별 어휘 지식의 양을 측정하였고 동시에 Gyllstad(2007)가 개발한 동사+명사 구조의 다어휘 표현 평가지인 ‘COLLMATCH’를 모델로 하여 다어휘 표현의 이해능력(receptive skill)과 표현능력(productive skill)을 측정하였다. 그 결과 개별 어휘 지식이 많은 학습자일수록 다어휘 표현의 평가에서도 더 높은 점수를 받았다. 이는 분명 개별 어휘 지식이 다어휘 표현 지식에도 중대한 영향을 미친다는 것을 의미한다. 하지만 다어휘 표현의 이해 능력 평가 점수와 표현 능력 평가 점수 사이에는 큰 차이를 보였다. Mutlu와 Kaşlıoğlu는 이러한 결과를 바탕으로 다어휘 표현을 실제 사용할 수 있는 연습 기회를 충분히 제공할 필요가 있다고 제안하였다. Chui(2006)는 홍콩 대학생들을 대상으로 Coxhead(2000)의 Academic Word List(AWL)에 대하여 모국어로 의미를 듣고 그에 해당되는 학술 어휘를 영어로 말하는 ‘Productive Vocabulary Levels Test (PVLT)’을 적용하여 학술 어휘 지식을 측정하였다. 이후 개별 어휘의 질적 평가(vocabulary depth test) 중 한 영역으로 다어휘 표현을 동일한 방식으로 측정하였다. 그 결과 두 평가 결과 간의 상관계수는 .69로 나타났다. Shimamoto(2000)는 모국어로 된 의미를 보고 그에 해당하는 목표어 개별 어휘를 선택지에서 찾는 ‘Vocabulary Levels Test(VLT)’를 활용하여 개별 어휘 지식을 측정한 후 같은 방식으로 다어휘 표현의 지식을 측정하였다. Shimamoto의 연구 결과에서는 두 가지 지식의 상관계수가 .73으로 나타났다. 또한 Koizumi(2005)의 연구에서도 개별 어휘 지식과 다어휘 표현 지식 간에는 .67로 전체적으로 보면 약 .70 내외의 상관도를 보이는 것으로 나타났다.

여기서 기존 연구들의 공통적인 문제점을 살펴보면 일부 연구에서는 절대 개별 어휘 지식을 측정할 수 있는 개별 어휘 능숙도 평가지(예, VST)를 적용하였지만 대부분의 경우는 그렇지 못했다. 또한 다어휘 표현 평가지의 경우에도 임의로 선정된 다어휘 표현만 측정하여 절대적인 다어휘 표현 지식을 측정하지 못하는 한계를 보였다.

III. 연구 방법

1. 데이터 수집

본 연구에서 활용한 한국인 학습자 구어 코퍼스와 문어 코퍼스 구축에는 서울 소재 대학교의 교양영어 과목을 수강한 134명의 대학생들이 참여하였고 이 중 여학생은 101명이며 남학생은 33명이었다. 또한 93명은 1학년, 27명은 2학년, 14명은 3학년에 재학 중이었으며 이들이 수강한 ‘Reading and Discussion’이란 과목은 전공에 상관없이 졸업을 위해 모든 재학생이 수강해야 하는 필수 과목이었기 때문에 참여자들의 전공은 다양했다. 또한 대부분의 학생들은 영어권 국가에서 장기 체류한 경험이 없었다.

구어 자료는 짧은 개별 인터뷰를 통해 수집되었으며 인터뷰에서는 자기소개, 취미나 관심사, 앞으로의 희망 등 다양한 질문들을 활용하였다. 문어 자료 또한 인터뷰 주제 가운데서 자유롭게 선택하여 250~350단어 길이의 에세이를 작성하는 방식으로 수집되었다. 다음의 표 1은 한국인 학습자의 구어 및 문어 코퍼스의 규모를 비교한 것이다.

표 1
한국인 대학생 구어 및 문어 코퍼스 비교

구분	출현형(Token)	개별 어휘 유형(Type)
구어 코퍼스	33,384	2,317
문어 코퍼스	30,675	2,972

구어 코퍼스는 33,384개의 단어(token)로 구성되었으며 문어 코퍼스는 30,675개의 단어 구성되어 두 코퍼스의 규모의 약 30,000단어로 거의 유사했다. 또한 개별 어휘의 유형(type)도 구어에서 2,317개로 나타났고 문어에서는 2,972개로 나타나 큰 차이를 보이진 않았다. 한국인 학습자 코퍼스와 비교할 영어 원어민 코퍼스로는 각 100만 단어로 구성된 BNC Spoken Sampler와 BNC Written Sampler를 사용하였다.

2. 분석 도구

1) 개별 어휘 분석 도구

본 연구에서는 코퍼스에 사용된 개별 어휘 사용의 양상을 분석하기 위해 Heatley와 Nation(2002)이 개발하여 무료 배포하고 있는 RANGE program에 등급별 어휘군(word family) 목록인 BNC-COCA 25000(Nation & Webb, 2011)을 탑재한 BNC-COCA 25000 RANGE program을 적용하였다. BNC-COCA 25000는 25개의 등급으로 구성되어 있으며 한 개 등급은 1,000개의 어휘군(파생과 굴절 변이형 포함)으로 구성되어 있다. 기초 핵심 개별 어휘인 상위 2,000개의 어휘 선정에는 구어 자료나 일상적인 주제가 주를 이루는 문어 자료 그리고 아동 문학 자료 등으로 구성된 별도의 코퍼스를 활용하였고 3,000-25,000 수준의 어휘군 즉 총 23,000개의 어휘군 선정에는 영국 영어를 대표하는 코퍼스인 BNC와 미국 영어를 대표하는 코퍼스인 Corpus of Contemporary American English(COCA)를 통합한 대규모의 언어 자료를 활용하였다. BNC-COCA 25000은 지금까지 개발된 등급별 어휘군 목록 가운데 가장 큰 규모이며 가장 일반적으로 개별 어휘 분석에 사용된다는 점을 고려하여 본 연구의 개별 어휘 분석에 적용하게 되었다.

BNC-COCA 25000 RANGE program의 사용 방법은 먼저 아래의 그림 1 상단에 있는 'File' 버튼을 클릭하여 분석하고자 하는 모든 코퍼스를 업로드 한다. 그 다음 4개의 코퍼스를 한 번에 각각 분석하기 위해 'BatchFiles'를 선택한다.

RANGE program의 기본설정은 3개의 개별 어휘 목록을 기반으로 하고 있지만 BNC-COCA 25000의 경우는 25개의 개별 어휘 목록으로 구성되어 있어 'Number of Baseword Files'에 25를 직접 입력해야 한다. 하지만 본 연구에서는 다어휘 표현의 분석 도구와 동일한 조건으로 개별 어휘 등급을 분석하기 위해 그림 1의 화면에서 보는 바와 같이 25개 목록 가운데 20개만 적용하였다. 설정이 완료되면 'Process Files'를 클릭하여 분석을 수행한다. 분석결과는 등급별 출현형(token) 수 및 비율, 개별 어휘 유형(type) 수 및 비율, 어휘군(family) 수의 기술 통계와 등급별로 사용된 구체적인 개별 어휘 목록을 제공한다.

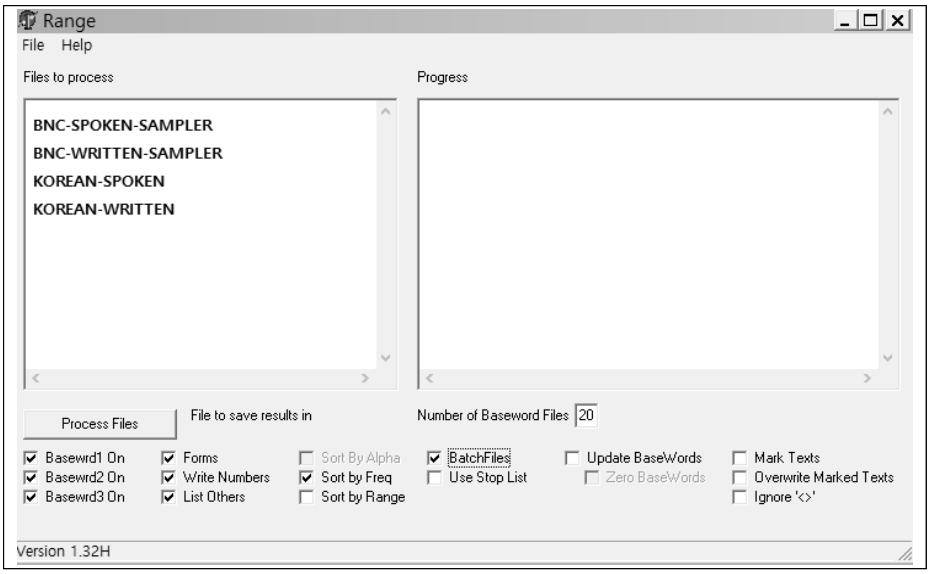

그림 1
BNC-COCA RANGE program의 사용자 인터페이스

2) 다어휘 표현 분석 도구

최근까지 다어휘 표현을 특정 다어휘 표현 목록에 기반하여 분석할 수 있는 프로그램은 존재하지 않았다. 하지만 Shin, Chon, Lee와 Park(2018)이 개발하여 본 연구에서 활용한 COCA_MWU20 ColloGram은 COCA에서 추출한 10,000개의 다어휘 표현군(multi-word unit family)이 탑재되어 있어 이를 바탕으로 다어휘 표현의 사용 양상을 RANGE program과 동일한 방식으로 분석할 수 있다. 일반적인 개별 어휘 목록의 등급이 1,000개 어휘군을 단위로 등급화되는 것과는 달리 20개 등급으로 구성된 COCA_MWU20은 다어휘군 표현의 빈도가 개별 어휘보다는 낮다는 점을 고려하여 등급을 세분화하였고 그 결과 각 등급은 500개의 다어휘 표현군을 포함하고 있다.

지금까지 다어휘 표현에는 소위 'family'라는 개념이 도입되지 않았으나

COCA_MWU20에서는 ‘다어휘 표현군’의 개념을 처음 적용하여 대표형에 어휘가 추가되거나 삭제되는 경우 이를 파생 변이형으로 간주하고 동사의 굴절 변화형(예, go home, goes home, going home, went home, gone home)과 명사의 단복수(예, year old, years old)를 포함하여 이를 굴절 변이형으로 정의하였다. 또한 ‘this moment’와 ‘at this moment’와 같이 다어휘 표현 유형별 빈도 계산에 중복이 있는 경우 ‘Subtractive(-) Method’를 적용하였다. 즉 ‘this moment’의 빈도 단순하게 산출하면 ‘at this moment’의 한 부분으로 포함된 ‘this moment’를 총 빈도에 중복으로 포함하되 때문에 순수한 'this moment'의 빈도를 산출하기 위해서는 ‘this moment’의 총 빈도에서 'at this moment'의 빈도를 제외할 필요가 있다. 기존의 다어휘 표현 분석 연구에서는 이와 같은 중복된 빈도 산출의 문제를 해결하지 못했고 또한 ‘go home’과 ‘went home’과 같은 유형을 별개의 표현으로 간주하는 한계가 있었다. 하지만 COCA_MWU20 ColloGram을 활용하면 이러한 문제를 모두 해결할 수 있다는 장점이 있다.

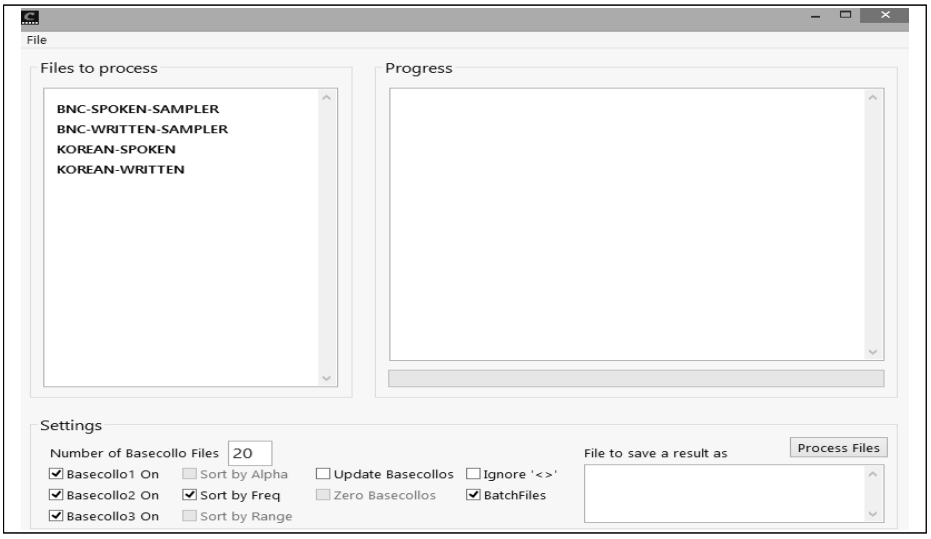

그림 2
COCA_MWU20 ColloGram의 사용자 인터페이스

COCA_MWU20 ColloGram은 사용 방법은 RANGE program과 동일하다. 다만 COCA_MWU20 ColloGram에서는 분석할 코퍼스의 업로드 시 사용자의 편의성을 고려하여 파일을 직접 박스 안에 집어넣는 드래그 앤 드롭(Drag & Drop) 방식이 추가되었다. 분석결과 또한 RANGE program과 마찬가지로 등급별 출현형(token) 수 및 비율, 다어휘 표현 유형(MWU type) 수 및 비율, 다어휘 표현군(MWU family) 수 및 비율의 기술 통계와 등급별로 사용된 구체적인 다어휘 표현의 목록을 제공한다.

IV. 연구 결과

1. 개별 어휘 및 다어휘 표현의 빈도 비교

표 1에서도 살펴 본 바와 같이 약 30,000단어로 구성된 한국인 학습자 구어 코퍼스의 개별 어휘 유형 수는 2,317개 그리고 문어 코퍼스는 2,972개로 차이가 크진 않지만 기존 연구와 마찬가지로 문어 자료에서 더 다양한 개별 어휘가 사용된다는 사실을 확인할 수 있었다. 또한 100만 단어의 BNC Spoken Sampler의 개별 어휘 유형 수는 18,127개, BNC Written Sampler는 43,038개로 문어 자료에서 약 두 배 이상의 다양한 개별 어휘 유형이 사용되어 한국인 학습자 코퍼스에 비해 보다 확연한 차이를 보였다.

다어휘 표현의 빈도 분석의 경우, 한국인 학습자 구어 코퍼스에서는 총 1,246개의 다어휘 표현이 사용되었고 문어 코퍼스에서는 1,084개로 나타나 문헌연구에서 살펴본 바와 같이 구어 자료에서 다어휘 표현이 보다 빈번하게 사용된다는 것을 알 수 있었다. 다어휘 표현 유형의 비교에서는 구어에서 486개, 문어에서 578개로 문어에서 보다 다양한 표현이 사용된 반면 구어에서는 특정 유형의 다어휘 표현이 보다 반복적으로 사용된다는 것을 알 수 있었다. BNC Spoken Sampler에서는 총 17,716개의 다어휘 표현의 사용이 확인되었고 그 유형의 수는 4,305개였다. 하지만 BNC Written Sampler에서는 총 다어휘 표현의 수가 20,407개였고 유형의 수도 7,520개로 문어 코퍼스에서 보다 빈번하고 다양하게 사용되어 기존의 연구와는 다소 차이가 있었다. 본 연구에서는 보다 정확한 분석을 위해 개별 어휘 및 다어휘 표현의 등급별 분포를 추가로 살펴보았다. 다음의 표 2는 각 코퍼스별 상위 5개 등급에 나타난 개별 어휘 및 다어휘 표현의 사용 분포 비율을 정리한 것이다.

표 2
코퍼스별 개별 어휘 및 다어휘 표현의 빈도수 분포 비율 비교

구분	한국인 학습자 코퍼스				원어민 코퍼스			
	개별 어휘		다어휘 표현		개별 어휘		다어휘 표현	
	구어	문어	구어	문어	구어	문어	구어	문어
Band 1	87.69	87.33	87.3	72.96	49.12	46.68	56.58	38.14
Band 2	5.25	5.24	3.63	9.13	15.57	11.99	13.29	11.97
Band 3	1.7	1.76	1.31	5.45	6.58	7.75	5.8	8.51
Band 4	0.43	0.54	0.66	1.86	5.46	4.43	3.82	5.62
Band 5	0.21	0.44	0.48	1.03	1.77	3.78	2.92	4.73
계	95.28	95.31	93.38	90.43	78.5	74.63	82.41	68.97

상위 5개 등급에 포함된 개별 어휘나 다어휘 표현은 전체 사용의 과반 이상

을 차지한다. 특히 개별 어휘의 경우는 상위 5,000개의 개별 어휘에 포함된 항목의 사용 비율이 모든 코퍼스에서 90% 이상을 보였다. 다어휘 표현은 상위 5개 등급, 즉 2,500개의 다어휘 표현에 포함되는 비율이 69~79%에 달했다. 공통적인 특징은 비원어민인 한국인 학습자들의 경우 구어나 문어에서 개별 어휘 및 다어휘 표현의 사용 비율의 차이가 크지 않은 반면 원어민의 경우는 다소 차이가 있었다. 또한 다어휘 표현 분석에서 원어민 문어 코퍼스의 경우 그 비율이 가장 낮아 원어민은 문어 모드에서는 몇 개 등급에 집중되기 보다는 상대적으로 다양한 수준의 다어휘 표현을 폭넓게 구사하는 것으로 나타났다.

본 연구에서는 실제 표현을 비교하기 위해 다음의 표 3과 4와 같이 코퍼스별 최상위 빈도 10개의 개별 어휘 및 다어휘 표현을 비교·분석하였다.

표 3
코퍼스별 최상위 빈도 10개의 개별 어휘 비교(어휘군 기준)

빈도별 순위	한국인 학습자 코퍼스의 개별 어휘		원어민 코퍼스의 개별 어휘	
	구어	문어	구어	문어
1	***I***	***I***	the	the
2	to	to	***I***	of
3	and	and	***you***	and
4	so	*the*	it	to
5	in	in	and	a
6	it	a	that	in
7	a	***that***	to	***for***
8	***want***	so	a	that
9	*the*	of	of	it
10	***but***	***we***	in	***on***

표 4
코퍼스별 최상위 빈도 10개의 다어휘 표현 비교(다어휘 표현군 기준)

빈도별 순위	한국인 학습자 코퍼스의 다어휘 표현		원어민 코퍼스의 다어휘 표현	
	구어	문어	구어	문어
1	***high school***	***high school***	***as well***	***as well***
2	close friend	most memorable	*come on*	*prime minister*
3	***years old***	***very hard***	***come in***	in order
4	speak English	***middle school***	go out	set up
5	***middle school***	***years old***	come back	years ago
6	very good	years ago	come out	carry out
7	***very hard***	come true	very good	years old
8	very important	for a long time	go back to	***come in***
9	listen to music	very happy	very much	so far
10	best friend	so happy	come up with	so much

위의 표에서는 최상위 빈도 10개의 개별 어휘 및 다어휘 표현의 항목만을 비교했지만 각 코퍼스별 특징을 어느 정도 파악할 수 있었다. 먼저 한국인 학습자 코퍼스의 구어와 문어 비교에서는 10개의 최상위 빈도 개별 어휘 중 7개 일치가 일치하였다. 개별 어휘는 구어와 문어 상관없이 기본적으로 기능어들이 최상위 빈도를 차지하기 때문에 상위 빈도 10개의 개별 어휘만을 비교했을 때 높은 일치도를 보였다. 한 가지 흥미로운 발견은 'I'라는 1인칭 대명사의 빈도가 구어와 문어 모두에서 가장 높았다는 것이다. 즉 한국인 학습자는 영어 원어민 화자에 비해 문어에서도 구어 스타일을 빈번하게 구사한다는 것을 알 수 있다. 한국인 학습자들이 구어에서 'want'란 동사를 매우 선호하는 것도 분석에서 확인하였고 정관사 'the'의 순위가 구어에서 상대적으로 낮은 것은 비원어민 화자가 원어민 화자에 비해 구어에서 관사를 신경 쓰지 못하는 경우가 더 많기 때문이라고 판단된다. 실제 다음의 표 5에 제시된 예문만 보아도 한국인 학습자들이 구어와 문어 스타일을 구분하여 사용하는데 미숙하다는 것을 알 수 있다.

표 5
한국인 학습자 문어 코퍼스에 나타난 구어 스타일의 표현 예시

-Form [sic] now on, I am gonna writing about my saddest experience.
-About two month [sic] ago, I had [sic] final test 수능.
-I want [sic] be a high-school Korean teacher.
-When I was eighteen, I have [sic] two best friends.

* [sic]은 오류가 있는 표현을 그대로 발췌한 것을 의미함

BNC Sampler의 개별 어휘 사용 또한 8개 항목이 일치하여 상위 빈도에서는 구어와 문어의 차이가 크지 않다는 것을 알 수 있었다. 차이를 보인 두 개 항목인 'I'와 'you' 1, 2인칭 대명사는 대화상에서 흔히 나타날 수 개별 어휘로 구어의 특성을 잘 보여주는 개별 어휘였다.

개별 어휘 분석결과와는 달리 상위빈도 10개의 다어휘 표현의 분석에서는 비원어민과 원어민의 코퍼스 모두에서 구어와 문어의 일치도가 매우 낮았다. 한국인 학습자 코퍼스에서 구어와 문어 모두에서 나타난 다어휘 표현은 4개였고 BNC Sampler의 경우에는 2개에 불과했다. 한국인 학습자 코퍼스에 나타난 상위 빈도 다어휘 표현은 데이터 수집에 활용된 인터뷰와 쓰기 주제의 영향이 큰 것으로 보이며 상대적으로 데이터 규모가 큰 BNC Sampler는 보다 보편적인 표현들이 순위에 올랐다. BNC Sampler의 구어와 문어에서 최고 빈도를 보인 'as well'의 경우 한국인 학습자 코퍼스에서는 순위에 들지 못했는데 그 이유는 한국인 학습자들은 '또한'이라는 의미의 표현으로 'as well'보다는 'too'를 선호하기 때문이라고 판단된다. 또한 BNC Sampler의 구어 자료에는 특별한 의미가 없는 간투사(interjection)로도 자주 쓰이는 'come on'이 포함되어 있었고(예, ~ Peter! *Come on*, man. There. Nearly finished.~) 문어 자료에서는 특이하게도 'prime minister'라는 다어휘 표현이 높은 순위를 보였다. 'prime minister'는 미국

코퍼스인 COCA에서도 높은 빈도를 보였다. 그 이유는 문어 코퍼스 구축에 흔히 포함되는 신문과 같은 시사적인 자료들의 특성이 반영된 것이라고 판단된다.

2. 개별 어휘 및 다어휘 표현 사용의 다양성 비교

본 절에서는 각 코퍼스별 개별 어휘 및 다어휘 표현 사용에 대해 전체적인 TTR 수치는 물론 등급별 TTR 수치를 비교하여 개별 어휘 및 다어휘 표현 사용의 다양성을 살펴보았다.

다음의 그림 3의 TTR 비교를 보면 다어휘 표현의 TTR 수치가 개별 어휘에 비해 월등히 높다. 이는 TTR이 비율로 계산하기 때문에 전체 모수(전체 사용 빈도수)가 작은 다어휘 표현에서 TTR 수치가 상대적으로 높게 나타났다. 즉 모수가 크면 표현이 반복적으로 사용될 가능성이 높아 다양성은 낮아지는 경향을 보인다. 따라서 사용 빈도수 면에서 큰 차이가 있는 개별 어휘와 다어휘 표현을 비교하기 보다는 이를 개별 어휘와 다어휘 표현으로 구분하여 분석하는 것이 보다 의미가 있다고 볼 수 있다.

개별 어휘의 TTR 수치를 보면 모수가 작은 한국인 학습자 코퍼스에서의 수치가 원어민 코퍼스의 수치보다 높게 나왔고 여기서 주목할 것은 한국인 학습자와 원어민 모두 문어 코퍼스에서 더 높은 수치를 보였다는 점이다. 이는 구어에 비해 문어에서 보다 다양한 표현이 사용되고 있다는 것을 말해준다. 수치 단위는 다르지만 이는 다어휘 표현에서도 동일하게 나타났다. 다어휘 표현에서도 한국인 학습자와 원어민 모두 문어 코퍼스에서 더 높은 수치를 보였다.

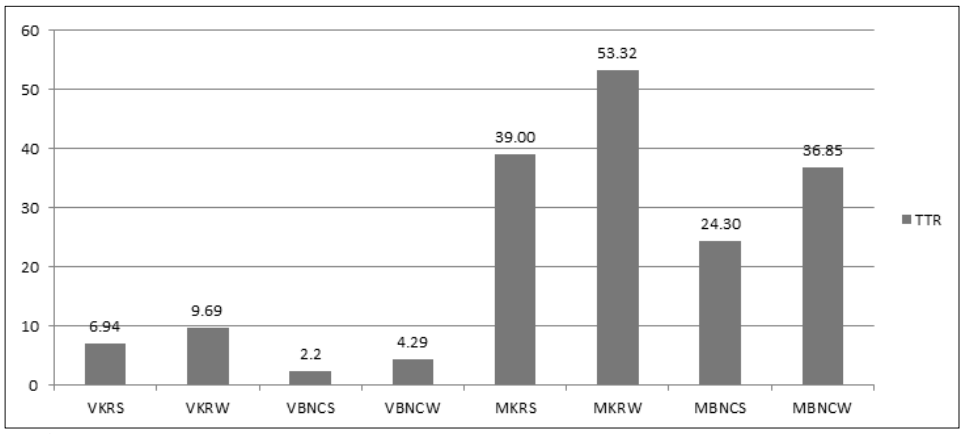

그림 3
코퍼스별 개별 어휘 및 다어휘 표현의 TTR 수치 비교[1]

[1] 약어 중 맨 앞의 V는 개별 어휘, M은 다어휘 표현을 의미하며 KR은 한국인 학습자 코퍼스, BNC는 영어 원어민 코퍼스, 맨 뒤의 S는 구어, W는 문어를 의미함(이후

본 연구에서는 한국인 학습자와 원어민 화자 간의 개별 어휘 및 다어휘 표현 사용의 다양성을 보다 심층적으로 분석하기 위해 등급별로 TTR을 재분석하였다. 앞서 표 2에서 살펴본 바와 같이 아래 그림 4를 보면 등급별 개별 어휘와 다어휘 표현의 사용이 최상위 빈도의 등급(Band 1)에 집중되고 있는 것을 확인할 수 있다.

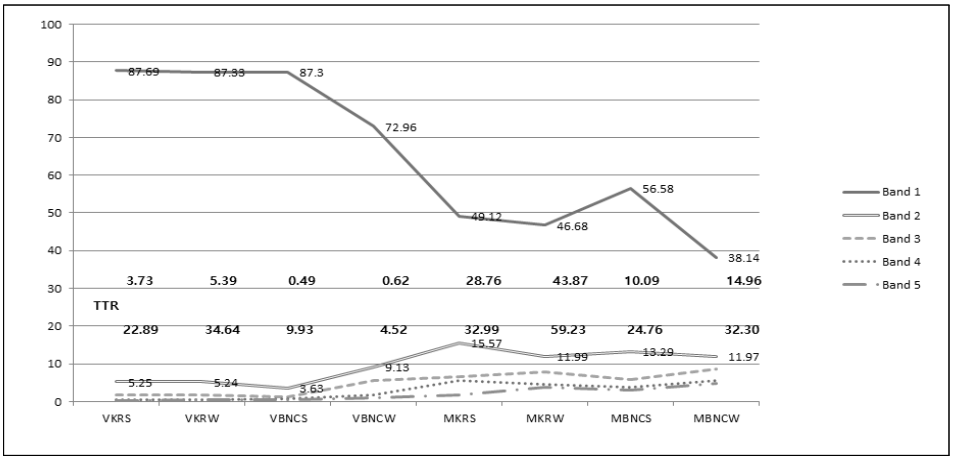

그림 4
코퍼스별 개별 어휘 및 다어휘 표현의 사용 비율에 따른 TTR 수치 변화

한국인 학습자의 구어 코퍼스에서는 최상위 빈도의 개별 어휘 등급(Band 1)에서 TTR 수치가 3.73인 반면 다어휘 표현은 28.76으로 나타나 모수가 클수록 TTR의 수치는 낮아진다는 사실을 다시 확인할 수 있었다. 이를 각 개별 어휘 및 다어휘 표현 등급에 적용한다면 등급이 낮아질수록 개별 어휘나 다어휘 표현의 총 사용 빈도수는 줄어들 것이고 이에 따라 개별 어휘 및 다어휘 표현 사용의 다양성은 상대적으로 높게 나타날 것이라는 것을 예상할 수 있다. 하지만 한국인 학습자 코퍼스에서는 그 예상에서 벗어난 결과를 보여주었다.

다음의 그림 5와 6을 보면 원어민의 경우 개별 어휘(VBNCS, VBNCW)나 다어휘 표현(NBNCS, MBNCW)의 TTR이 구어와 문어에 상관없이 등급의 하락할수록(Band 1→Band 10) 일정한 상승 패턴을 보이는 반면 한국인 학습자는 개별 어휘나 다어휘 표현, 구어나 문어에 상관없이(VKRS, VKRW, MKRS, MKRW) 매우 불규칙한 패턴을 보이고 있다. 이는 한국인 학습자가 구사하는 개별 어휘나 다어휘 표현 지식이 원어민이 구사하는 개별 어휘나 다어휘 표현 지식과 달리 등급별로 불균형을 이루고 있다는 것을 의미한다.

모든 그래프에 동일하게 적용됨)

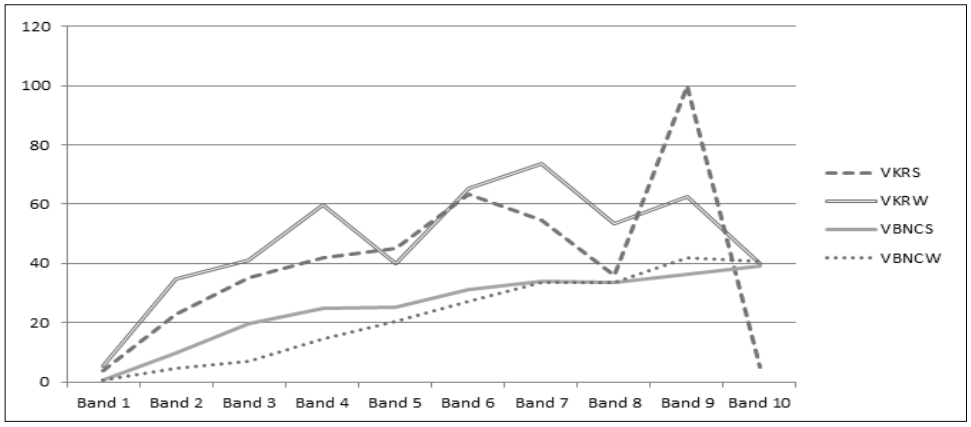

그림 5
등급별 개별 어휘의 사용 비율에 따른 TTR 수치 변화 비교

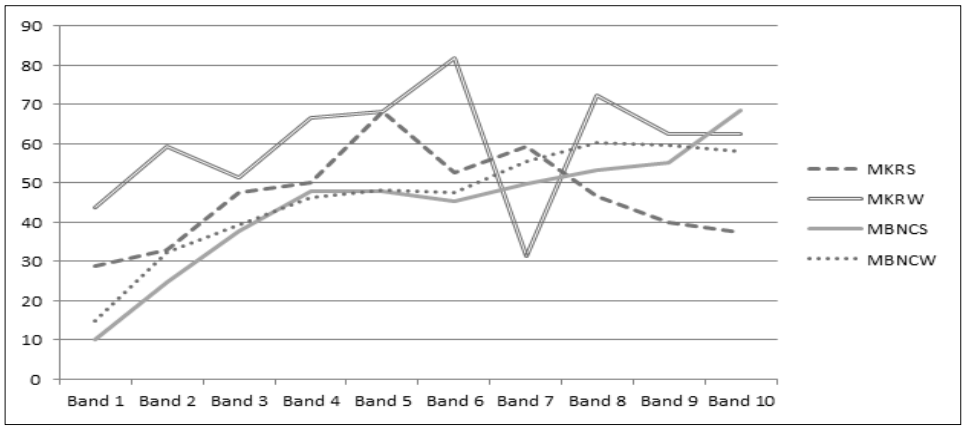

그림 6
등급별 다어휘 표현의 사용 비율에 따른 TTR 수치 변화 비교

실제 EFL 환경에서는 자연스런 언어 노출과 같은 우연적 학습(incidental learning) 통해 습득할 수 있는 개별 어휘나 다어휘 표현은 매우 제한적이며 대부분은 단어장 활용과 같은 의도적 학습(intentional learning)을 통해 개별 어휘 및 다어휘 표현 지식을 늘려나가는 것이 일반적이다. 하지만 이러한 의도적 학습에서 제공되는 개별 어휘나 다어휘 표현의 종류와 빈도는 원어민에게 자연스럽게 주어지는 등급별 노출 빈도와는 달리 입시 등과 같은 특수한 시험환경에서 빈번하게 사용된 표현에 집중되어 어려운 표현은 상대적으로 많이 알고 오히려 쉬운 표현을 모르는 경우가 많다(Shin, Chon, & Kim, 2011). 같은 맥락에서 위의 등급별 TTR 수치의 등락 또한 같은 맥락에서 우리나라의 특수한 환경적 요인이 영향을 미친 결과라고 판단된다.

V. 결론 및 제언

본 연구는 절대적인 개별 어휘 및 다어휘 표현 등급을 분석할 수 있는 BNC-COCA 25000 RANGE program과 COCA_MWU20 ColloGram을 활용하여 한국인 학습자와 영어 원어민의 구어 및 문어 코퍼스에 나타난 개별 어휘 사용과 다어휘 표현 구사를 다각도로 비교하고자 하였다. 본 연구의 주요 연구결과는 다음과 같다.

첫째, 한국인 학습자는 영어 원어민에 비해 구어와 문어에서 사용하는 개별 어휘 및 다어휘 표현에 차이가 상대적으로 적었다. 이는 언어 능숙도의 한계로 문어에서도 구어 스타일의 표현을 구사하기 때문이라고 판단된다.

둘째, 90% 이상의 개별 어휘는 상위 빈도 5,000(Bands 1-5)단어 그리고 다어휘 표현의 70% 이상은 상위 빈도 2,500개(Bands 1-5)의 표현에 집중되었다. 이 중에서도 500개의 다어휘 표현군으로 구성된 최상위 등급(Band 1)은 나머지 19개 등급과 비교할 때 월등히 높은 사용 빈도를 보여 학습의 선택 집중의 필요성을 제시해 주었다.

셋째, 기존 연구와 마찬가지로 구어에서는 소수의 표현이 반복적으로 사용되는 경향이 있고 문어는 상대적으로 다양한 표현이 사용되었다. 특히 본 연구에서 다어휘 표현은 구어에서의 사용이 두드러지는 것을 볼 수 있는데 이는 즉각적 발화를 위해 구어에서 요구되는 인지적 부담이 다어휘 표현에 대한 의존도를 높이는 주요 요인이라는 것을 다시 한 번 확인할 수 있는 결과이다.

넷째, 등급별 개별 어휘 및 다어휘 표현의 사용 분석은 한국인 학습자들이 원어민의 자연스런 노출 빈도와는 매우 다른 노출 기회를 가진다는 것을 알 수 있었다. 이는 전체적인 언어 노출의 양 뿐만 아니라 등급별 노출의 분포에서도 불균형을 이루고 있다는 것을 의미한다. 결과적으로 이를 통해 한국인 학습자는 보다 빈도가 높은 유용한 표현 대신 빈도가 낮고 유용성이 떨어지는 표현을 학습하는 데 시간과 노력을 쏟는 경향이 있다는 것을 확인할 수 있었다. 앞서 1925년 Harold Palmer의 인터뷰에서도 소개하였지만 본 연구의 결과를 보면 안타깝게도 21세기 한국인 학습자들은 여전히 1920년대 일본인 학습자들의 모습과 닮아있다. 본 연구에서는 이러한 연구결과를 바탕으로 본 연구에서는 다음과 같은 시사점을 도출하였다.

먼저 학습의 효율성 제고를 위해서는 개별 어휘나 다어휘 표현의 등급을 고려하여 자연스러운 노출 빈도에 가깝게 등급별 개별 어휘 및 다어휘 표현을 교재 개발에 체계적으로 반영하고 개별 어휘나 다어휘 표현의 명시적 교육에서는 BNC-COCA 25000이나 COCA_MWU20과 같은 개별 어휘나 다어휘 표현 목록을 참고하여 학습내용을 설정할 필요가 있다. 또한 기존 연구(Laufer, 1992; Read, 2004)와 본 연구에서 보듯 개별 어휘의 경우는 상위 빈도 5,000개(Bands 1-5)에 집중 할 필요가 있으며 다어휘 표현은 Erman과 Warren(2000)의 연구에서 언급한 바와 같이 52-59%, 즉 전체 개별 어휘 사용의 약 절반가량이 다어휘 표현이

라는 분석에 비추어 상위 빈도 2,500개(Bands 1-5) 표현을 집중적으로 학습할 필요가 있다. 그리고 기초 수준부터 구어와 문어의 차이를 강조할 필요는 없지만 의사소통이 가능한 수준부터는 구어와 문어 스타일의 차이를 구분하여 지도할 필요가 있다.

끝으로, 본 연구에서는 기존 연구들에서는 시도한 바 없는 다어휘 표현의 TTR을 ColloGram이라는 프로그램을 활용하여 분석하였다. 하지만 이러한 새로운 시도에도 불구하고 일반적으로 코퍼스의 크기가 커질수록 TTR의 수치는 낮아지기 때문에 보다 정확한 어휘의 다양성을 측정하기 위해서는 STTR을 적용하는 것이 최근의 추세이다. 그럼에도 불구하고 다어휘 표현의 경우 개별 어휘에 적용되는 STTR의 산출 방식을 그대로 적용하는 것은 보다 신중을 기할 필요가 있다. Ha, Sicilia-Garcia, Ming과 Smith(2002)는 'n-gram'의 길이에 따라 Zipf's law를 달리 적용하는 연구를 발표한 바 있다. 보통 빈도수(f)×순위(r)=일정한 수치(C)라는 공식을 코퍼스를 구성하는 개별 어휘에 적용할 수 있지만 이들의 연구에서는 두 단어로 구성된 다어휘 표현(bi-gram)에는 $f * r^{.65}=C$를 적용한 한 바 있다. 따라서 STTR의 경우에도 일괄적으로 1,000개의 단어 수 당 나타나는 다어휘 표현의 TTR을 평균으로 산출할 것이 아니라 다어휘 표현을 구성하는 구성소의 수에 따라 STTR을 산출하는 공식이 새롭게 고안될 필요가 있다고 판단된다. 이러한 이유로 본 연구의 제한점으로 밝힌 다어휘 표현의 STTR에 대해서는 후속 연구를 통해 심층적으로 다시 연구될 필요가 있다.

참고문헌

김혜선, 윤현숙. (2010). 명시적·암시적 연어 지도가 학습자들의 수용적·표현적 연어 지식 습득에 미치는 영향. *영어영문학 연구, 52*(1), 129-149.

신동광. (2015). 어휘학습을 위한 언어 입력의 제공원으로서 원어민, 비원어민 교사, 교재 비교 연구. *외국어교육, 22*(4), 181-200.

유경미, 김낙복. (2014). 연어를 활용한 어휘지도가 초등학생의 영어 읽기 및 쓰기 능력에 미치는 영향. *영어교과교육, 13*(3), 83-104.

이재근, 김은주. (2008). 연어(collocation)를 활용한 초등학교 영어 쓰기 향상 방안. *영어교과교육, 7*(1), 83-94.

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

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research, 10,* 245–261.

Bresnan, J. (1999, August). *Linguistic theory at the turn of the century.* Paper presented at the 12th World Congress of Applied Linguistics, Tokyo, Japan.

Chui, A. S. Y. (2006). A study of the English vocabulary knowledge of university students in Hong Kong. *Asian Journal of English Language Teaching, 16*, 1-23.

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly, 34*, 213-238.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse, 20*(1), 29-62.

Gyllstad, H. (2007). *Testing English collocations: Developing receptive tests for use with advanced Swedish learners.* Unpublished doctoral dissertation. Lund University, Stockholm, Switzerland.

Ha, L. Q., Sicilia-Garcia, E., Ming, J., & Smith, F. J. (2002). Extension of Zipf 's law to words and phrases. *Proceedings of the International Conference on Computational Linguistics, Taiwan, 1*, 1-6.

Halliday, M. A. K. (1985) *An introduction to functional grammar.* London: Edward Arnold.

Heatley, A., & Nation, I. S. P. (2002). *RANGE and FREQUENCY programs* [Computer Software]. Wellington, New Zealand: Victoria University of Wellington. Retrieved December 20, 2017, from the World Wide Web: http://www.victoria.ac.nz/lals/about/staff/publications/BNC_COCA_25000.zip.

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.

Ishikawa, S. (2015). A consideration of the difference between the spoken and written English of native speakers and Japanese learners: A corpus-based study. *Discourse and Interaction, 8*(1), 37-52.

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

Koizumi, R. (2005). *Relationships between productive vocabulary knowledge and speaking performance of Japanese learners of English at the novice level*. Unpublished doctoral dissertation: University of Tsukuba, Tsukuba, Japan.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In Arnaud, P. J. L. & Bejlint, H. (Eds.), *Vocabulary and applied linguistics* (pp. 126-132). London: Macmillan.

Lee, D. Y. (2001). Defining core vocabulary and tracking its distribution across spoken and written genres: Evidence of a gradience of variation from the British National Corpus. *Journal of English Linguistics, 29*(3), 250-278.

Leech, G., Rayson, P., Wilson, A. (2001) *Word frequencies in written and spoken*

English: Based on the British National Corpus. London: Routledge.

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

Liu, D. (2003). The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly, 37*(4), 671-700.

McEnery, T., & Hardie, A. (2012) *Corpus linguistics: Method, theory, and practice*. Cambridge: Cambridge University Press.

Mutlu, G. & Kaşlıoğlu, Ö. (2016). Vocabulary size and collocational knowledge of Turkish EFL learners. *Journal of Theory and Practice in Education, 12*(6), 1231-1252.

Nation, I. S. P. (Ed.). (2016). *Making and using word lists for language learning and testing.* Amsterdam: John Benjamins Publishing Co.

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

Nation, I. S. P., & Webb, S.(2011). *Researching and analyzing vocabulary*. Boston: Heinle Cengage Learning.

Palmer, H. E. (1999 [1925]). Conversation. In R. c. Smith(Ed.), *The writings of Harold E. Palmer: An overview* (pp. 185-191). Tokyo: Hon-no-Tomosha.

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics, 24*, 146-161.

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

Shin, D. (2007). The high frequency collocations of spoken and written *English. English Teaching, 62*(1), 199-218.

Shin, D., Chon, Y., & Kim, H. (2011). Receptive and productive vocabulary sizes of high school learners: What next for the basic word list? *English Teaching, 66*(3), 123-148.

Shin, D., Chon, Y., Lee, S., & Park, M. (2018). *COCA_MWU20 ColloGram* [Compute r Software]. Seoul, South Korea: e-future. Retrieved December 20, 2017, from the World Wide Web: http://cfile281.uf.daum.net/attach/99EBA0495A80F11030 4D74.

Shirato, J. & Stepleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research, 11*(4), 393-412.

적용가능 수준 (Applicable Levels): elementary, secondary, and tertiary

신동광
광주교육대학교 영어교육과
61204 광주광역시 북구 필문대로55
Tel: (062) 520-4211
Email: sdhera94@gnue.ac.kr

전유아
한양대학교 사범대학 영어교육과
04763 서울특별시 성동구 왕십리로 222
Tel: (02) 2220-1144
Email: vylee52@hanyang.ac.kr

이신웅
한양대학교 인문과학대학 영어영문학과
04763 서울특별시 성동구 왕십리로 222
Tel: (02) 2220-0745
Email: shinwoonglee@hanyang.ac.kr

박명수
상명대학교 글로벌인문학부대학 글로벌지역학부
31066 충남 천안시 동남구 상명대길 31
Tel: (041) 550-5147
Email: myongsu@smu.ac.kr

Received March 14, 2018
Revised April 25, 2018
Accepted May 12, 2018