http://dx.doi.org/10.5392/JKCA.2014.14.11.1019

한국 예비 대학생의 영어 사용 특성 파악을 위한 대규모 공개 영어 학습자 코퍼스 구축 및 분석

Compilation of the Yonsei English Learner Corpus (YELC) 2011 and Its Use for Understanding Current Usage of English by Korean Pre-university Students

이석재[*], 정채관[**]
연세대학교 영어영문학과[*], 한국교육과정평가원 영어교육센터[**]

Seok-Chae Rhee(scrhee@yonsei.ac.kr)[*], Chae Kwan Jung(ckjung@kice.re.kr)[**]

요약

최근 영어 학습자 코퍼스(English learner corpus)를 활용하여 다양한 영어 교육 분야에 활용하는 시도가 이뤄지고 있다. 하지만 지금까지 국내에서 개발된 대다수 영어 학습자 코퍼스는 소규모이거나 공개가 되지 않아 공익을 위한 영어 교육 콘텐츠로서의 적절한 역할을 하지 못하고 있다. 본 연구에서는 국내외 영어 학습자 코퍼스 구축 현황을 살펴보고 대규모 공개 한국인 영어 학습자 코퍼스의 필요성을 논의한다. 또한, 이와 같은 필요성을 바탕으로 1백만 단어 이상으로 만들어진 대규모 공개 한국인 영어 학습자 코퍼스 구축 과정과 결과를 분석하여 예비 대학생의 영어사용 특성을 파악하고 이를 영어 교육 개선을 위해 활용할 수 있는 방안을 제안한다.

■ 중심어 : | 영어 | 영어교육 | 영어교육 콘텐츠 | 코퍼스 | 영어 학습자 코퍼스 | 예비 대학생의 영어 사용 특성 |

Abstract

In recent years, researchers have become increasingly interested in the creation and pedagogical use of English learner corpora. Many studies have shown that learner corpora can not only make a significant contribution to second language acquisition research but also contribute to the construction and evaluation of language tests by advancing our understanding of English learners. So far, however, little attention has been paid to the Korean EFL (English as a foreign language) learners' corpus. The Yonsei English Learner Corpus (YELC 2011) is a specialized, monolingual, and synchronic Korean EFL learner corpus that was developed by Yonsei University from 2011 to 2012. Over 3,000 Korean high school graduates (or equivalents) who were accepted by Yonsei University for their further studies participated in this project. It consists of 6,572 written texts (1,085,828 words) at nine different English proficiency levels. In this paper, we describe its compilation, and more specifically, how we have corpusized from a text archive to a corpus. After introducing the process of corpusization, we report arresting insights into the specific linguistic features that different proficiency levels of Korean learners of English have. This study also discusses the potential use of the YELC 2011 which is now freely available for research purposes.

■ keyword : | English | English Education | Corpus | English Learner Corpus | Corpus Compilation |

접수일자 : 2014년 08월 12일
수정일자 : 2014년 09월 22일
심사완료일 : 2014년 09월 22일
교신저자 : 정채관, e-mail : ckjung@kice.re.kr

I. 서론

코퍼스(corpus)란, 기본적으로 텍스트화 된 언어의 모음을 의미한다. 하지만 Sinclair[1]은 이러한 텍스트가 단순히 텍스트가 아니라 우리가 자연스럽게 발화한 언어이어야 한다고 주장한다. 한편, O'Keeff, McCarthy와 Carter[2]는 이러한 언어 모음이 종이 형태가 아니라 컴퓨터를 통해 읽을 수 있는 전산화 작업이 이뤄진 형태여야 한다며, 컴퓨터 발명 이후의 코퍼스에 의미를 새롭게 부여한다. 또한, Hunston[3]은 코퍼스를 구성하는 텍스트는 문어(written language)일 수도 있고 구어(spoken language)일 수 있다는 내용을 추가하여 코퍼스의 의미를 보다 구체화하였다.

코퍼스는 사회언어학을 포함, 문학을 비롯한 이론과 실제를 연결하는 다양한 응용언어학 분야에서 사용되어왔고[4][5]. 우리가 실제로 사용하는 언어를 모아 구축하는 코퍼스의 유용성은 일찍이 역설되어왔지만, 대규모 코퍼스를 구축하는 일에는 막대한 비용, 시간, 노력, 전문 인력이 필요하다[6]. 따라서 국내에는 아직 연구목적으로 연구자들이 자유롭게 사용할 수 있는 대규모 개방형 코퍼스는 거의 전무한 상태이다.

이러한 점에 착안하여, 본 연구에서는 국내 한 4년제 대학교 입학예정자들이 생산한 텍스트를 기반으로 영어 학습자 3,286명이 기여한 대규모 공개 한국인 영어 문어 코퍼스인 연세 영어 학습자 코퍼스(Yonsei English Learner Corpus, YELC 2011)를 구축하고, 이를 기반으로 우리나라 예비 대학생의 전반적인 영어 사용 특성과 등급별 특성을 분석하고자 한다. 본 연구는 우리나라 최초의 대규모 공개 한국인 영어 학습자 코퍼스를 구축함으로써 대학에 입학하는 학생들의 영어 사용의 실태를 파악하고, 영어 원어민과 어떤 차이가 있는지 규명하고, 또한 영어 학습자 코퍼스를 공개함으로써 국내외 학자와 연구자들이 영어 학습자의 실물 자료를 바탕으로 폭넓게 영어 교육에 활용할 수 있는 토대를 마련하는데 그 의의가 있다.

II. 코퍼스 구축 동향

컴퓨터로 코퍼스 검색이 가능한 근대 코퍼스의 시초로 불리는 것은 미국에서 만들어진 브라운 코퍼스(Brown Corpus, BC)라 할 수 있다. BC는 미국 브라운 대학교에 재직 중이던 Henry Kucera와 Nelson Francis가 당시 미국에서 사용되고 있는 미국 영어를 분석하기 위해 만들었다[7]. 이들은 1961년 이후 미국에서 발행된 신문기사, 사설, 추리소설 등 다양한 장르의 텍스트 500개를 선별한 후, 장르마다 각 2천 단어 내외로 균형을 맞춰 1백만 단어로 이뤄진 대규모 코퍼스를 구축하였다. BC는 이후 근대 코퍼스 구축의 첫 사례로 꼽히며 50년이 지난 지금에도 세계 여러 연구자에 의해 널리 활용되고 있다[8][9].

미국영어 연구를 위한 코퍼스 구축에 자극받은 영국 학자들은 영국영어 연구에 대한 필요성을 절감하여, 1960년대 중반부터 1990년대 사이 영국인들이 사용한 영어를 모아 영국국립코퍼스(British National Corpus, BNC)를 구축했다[10]. BNC에는 신문기사, 학술교재, 소설, 자원봉사자의 대화, 라디오 인터뷰 등 총 4,124개의 텍스트(약 1억 단어)가 투입됐고, BC에 비해 다소 늦게 구축됐지만, 문어만 있던 BC와 달리 90%의 문어와 10%의 구어로 구성되었다[11]. 따라서 BNC는 BC의 1백배라는 규모, 문어와 구어를 모두 포함한다는 점에서 BC의 한계를 보완했다고 할 수 있다.

BC나 BNC가 영어 원어민의 영어를 수집했다면, 영어 학습자 코퍼스(English learner corpus)는 영어 학습자들의 실제 영어 사용을 수집한다[12]. 영어 학습자 코퍼스 구축이 이뤄진 배경에는 1990년대 초반부터 영어 교수·학습 교재가 영어를 사용하는 영어 원어민 지필진의 직관이나 인위적인 예문으로 사전이나 교재가 제작되기보다는, 영어 학습자들이 발화하는 실제 영어 사용에 근거를 두고 영어 사전이나 교재가 만들어져야 한다는 인식이 늘어났기 때문이다[13][14]. 이와 같은 영향으로 만들어진 대표적인 영어 학습자 코퍼스로는 Longman Learners' Corpus(LLC)와 Cambridge Learner Corpus(CLC), International Corpus of Learner English(ICLE)가 있다.

영어 학습자 코퍼스 연구는 영어 원어민의 시각에서 영어 학습자들이 어려움을 겪을 것이라는 짐작 대신, 영어 학습자들이 실제 쓰고 말한 영어 사용의 실례 기

반으로 연구가 이뤄진다. 따라서 연구자는 이러한 영어 학습자들의 실제적인 언어사용을 분석함으로써 영어원어민보다 과도하게 많이 사용하거나 지나치게 적게 사용하는 단어와 표현을 직관이 아니라 연구를 통해 규명할 수 있다[15]. 예를 들어 Gillard와 Gadsby[16]는 영어 학습자들이 'nice', 'happy', 'big'이라는 형용사를 과도하게 사용하고, 'enormous', 'massive', 'huge'와 같은 형용사는 영어원어민보다 상대적으로 적게 사용한다는 것을 영어 학습자 코퍼스 연구를 통해 밝혀냈다. 이들은 영어 학습자 코퍼스 연구를 통해 영어 학습자가 혼란스러워할 만한 약 1,200개의 단어를 영어사전에 도움 상자 형태로 포함해 학습자가 이들 단어에 대해 더욱 세심한 주의를 기울일 것을 제안했다.

2000년대 후반 국내에서도 영어 학습자 코퍼스의 필요성에 대한 인식이 늘어났고, 본격적으로 한국인 영어 학습자 코퍼스 구축이 시도되었다. (주)능률교육에서는 사전제작을 위해 The Neungyule Interlanguage Corpus of Korean Learners of English(NICKLE)이라는 영어 학습자 코퍼스를 구축하려 시도했다[17]. NICKLE은 대학생들이 작성한 문어와 구어로 구성되었고, 그 규모는 약 890,000단어로 알려져있다(http://www.uclouvain.be/en-cecl-lcworld.html). NICKLE이 우리나라 최초로 전국 단위 대학교에서 수집한 영어 학습자의 실물 언어자료를 바탕으로 영어 학습자 코퍼스를 구축하려 했다는 점은 의미하는 바가 크다. 하지만 NICKLE은 해당 과제가 회사 내부 사정상 잠정 중단됐고, 현재까지 그 데이터는 일반에 공개되지 않고 있다[18].

(주)청담러닝에서는 자사에서 운영 중인 어학원을 통해 영어 전치사 오류 교정 모델 개발을 위한 Chungdahm English Learner Corpus(CELC)를 만들었다. 한국어가 모국어인 10-16세 사이의 학생들이 작성한 작문을 수집해서 만들어진 CELC는 그 규모가 1억 3천만 단어에 달한다. 하지만 저자들도 인정했다시피 "이들 자료는 현재 엄밀한 의미의 정제된 코퍼스라기보다는 데이터베이스의 형태로 존재하고 있다"[19]. 다시 말해, CELC는 코퍼스 언어학에서 통용되는 '코퍼스'라기 보다는 학생들이 작성한 작문을 모아 놓은 문서 보관소(text archive)로 보는 것이 적절하다. CELC 역시 NICKLE와 마찬가지로 회사 내부 사정상 담당 부서가 없어졌고, 외부연구자의 사용이 불가능하다[20].

출판사와 사교육 업체의 주도로 코퍼스 구축이 시도된 경우를 제외하고, 국내에서 주목할 만한 영어 학습자 코퍼스는 The Seoul National University Korean-speaking English Learner Corpus(SKELC)를 예로들 수 있다. SKELC는 2005년부터 2008년까지 서울대학교 학생들의 작문을 모아 만든 한국인 영어 학습자 문어 코퍼스로 약 899,505단어로 이루어져 있다[21]. SKELC가 처음으로 대규모 한국인 영어 학습자 코퍼스의 지평을 열었다는 점에서는 큰 의미가 있지만, 서울대학교 내부적으로만 사용할 수 있다는 현실적인 제약이 있다.

이처럼 국외뿐 아니라 국내에서도 차츰 한국인 영어 학습자 코퍼스의 중요성과 필요성에 공감하고, 영어 학습자 코퍼스 구축에 대한 시도가 있었지만 예기치 못한 사정으로 중단되었거나, 엄밀한 의미에서 코퍼스화(corposization)로 되지 못한 상태이거나, 외부 연구자가 연구목적으로 자유롭게 사용할 수 없는 제한이 있다.

III. YELC 2011의 배경

1. YELC 2011의 텍스트 소스

현행 대학입학수학능력 영어 시험은 수험생의 읽기와 듣기능력 위주로 평가한다. 따라서 대학입학수학능력 영어 시험을 치르고 입학한 대학 신입생의 말하기와 쓰기 능력을 평가하기란 거의 불가능하다. 연세대학교는 자체 개발한 신입생 영어 능력 평가(Yonsei English Placement Test, YEPT)를 통해 2009년부터 예비 대학생의 말하기와 쓰기 능력을 진단해왔다. 영어 원어민이 다수를 차지하는 언더우드 국제대학에 입학하는 학생과 면제에 해당하는 공인 영어 시험 성적표를 제출한 학생을 제외하고 수시와 정시 각 입학전형에 관계없이 연세대학교 입학 예정자 전원이 YEPT에 응시해야 한다.

YEPT는 컴퓨터 기반 테스트를 통해 이뤄진다. 수험생은 시험이 시작되기에 앞서 본인 확인 과정을 거친

후, 시험 결과를 연구목적으로 사용할 수 있다는 것에 동의한다. 본격적으로 시험이 시작되면, 수험생은 컴퓨터 스크린을 통해 시험에 관한 일반적인 정보를 전달받은 후 마이크로폰이 부착된 헤드셋을 이용하여 컴퓨터 화면에 나오는 질문에 대답한다. 수험생이 말한 음성내용은 엠피스리(MP3) 파일 형태로 중앙 컴퓨터 서버에 실시간으로 저장된다. 쓰기 시험은 컴퓨터 화면에 질문이 나오면 전자사전이나 인터넷 검색 도움 없이 그 자리에서 바로 작문을 하는 방식으로 이뤄진다.

말하기 영역은 총 20분에 걸쳐 진행되며, 파트 1과 2(단답형으로 대답하기), 파트 3(짧은 지문을 소리 내어 읽기), 파트 4와 5(친숙한 주제에 관해 말하기), 파트 6, 7, 8(학술적인 주제에 관해 말하기) 등으로 이루어져 있다. 쓰기 영역은 60분에 걸쳐 진행되며, 총 세 부분으로 구성되어 있다. 파트 1에서는 주어진 단어를 올바른 순서대로 입력하기, 파트 2는 친숙한 주제에 대해 최대 100단어 내외로 작성하기(e.g., What was your favorite extracurricular activity in high school? What made you join the activity?), 파트 3에서는 학술적인 주제에 대해 최대 300단어 내외로 작문을 한다(e.g. Why should people receive a college education? State your opinion).

시험 결과는 모두 중앙 컴퓨터 서버에 실시간으로 저장되며, 평가자 훈련을 받은 영어 원어민 혹은 동등한 자격을 갖춘 평가자에 의해 채점된다. 이때 평가는 Common European Framework of Reference for Languages(CEFR)을 토대로, CEFR에서 제시한 6등급을 우리나라 대학입학수학능력 시험의 9등급에 맞춰 세분화시킨 9등급으로 등급이 매겨진다[22]. 해외에서 이같이 영어 학습자들의 시험 결과를 바탕으로 만들어진 유사한 코퍼스로는 Educational Testing Service(ETS)에서 구축한 TOEFL 2000 Spoken and Written Academic Language Corpus(T2K-SWAL), Japanese English as a Foreign Language Learner(JEFLL) Corpus 등이 있다[23][24].

2. YELC 2011의 대표성

코퍼스를 구축할 때 고려해야 할 중요한 요소 중 하나는 코퍼스의 대표성(representativeness)이다. 영국에서 구축한 British Academic Written English(BAWE)의 경우 3개 영국 대학교 학부과정 학생들의 과제물을 모은 것으로, BAWE를 이루는 텍스트의 50% 이상은 3개 대학교 중 1개의 특정 대학교에서 집중적으로 수집되었다[25][26]. 이와 같은 한계가 있음에도 BAWE의 대표성을 인정하는 이유는, 1) 30개 이상의 전공 2) 1,009명의 기여자 3) 6백5십만 단어 이상의 대규모로 요약될 수 있다[27]. 현실적으로 어떤 코퍼스도 목표 연구대상과 장르를 100% 대표할 수 없고, 이것은 코퍼스가 가진 태생적 한계이기도 하다. 따라서 현실을 반영하여 코퍼스는 '제한적' 대표성을 가질 수밖에 없다는 것이 저자들의 생각이다. BAWE 역시 이 같은 코퍼스로의 대표성을 세계 각지의 연구자들로부터 인정받았기 때문에 2008년 공개가 된 이후 활발한 연구자료로 활용되고 있다.

YELC 2011도 BAWE와 마찬가지로 국내 4년제 대학교 1개교에 입학할 예정자들이 전국에 있는 200개 이상의 4년제 대학 입학예정자 전부를 대변할 수는 없다. 하지만 YELC 2011는, 1) 수시와 정시 기여자(수시 모집: 80%, 정시 모집: 20%) 2) 전국단위 고등학교 출신 기여자 배경, 3) 다양한 국내 고등학교 유형을 대부분 포함, 4) 정원 내·외 수시(진리·자유 전형, 사회기여자 전형, 사회적 배려대상자 전형, 연세한마음 전형)과 정시(농어촌학생 특별전형, 특수교육대상자 특별전형, 전문계고교출신자 특별 전형, 연세한마음 전형, 새터민 특별전형) 포함, 5) 총 3,000명 이상에 달하는 예비 대학생을 대상으로 1백만 단어 이상의 대규모로 구성되어있다. 국내에서 상대적으로 우수학생들이 입학한다는 연세대학교 입학예정자들의 영어 실력이 다른 학교 학생들과 비교하면 높을 수도 있다는 직관적인 의견도 있을 수 있다. 하지만 영어 성적이 출중한 학생들은 YEPT를 이미 면제를 받은 상태이고, 영어 성적이 높지 않을 수도 있는 상당수 전형 학생들이 포함되었다는 전후 상황을 반영하면, 이 연구에서 YELC 2011이 한국인 영어 학습자이고 국내 4년제 대학교에 입학할 예정자의 영어 쓰기 능력을 보여주는 '제한적인' 대표성을 가진다고 할 수 있다고 판단한다. YELC 2011 역시 BAWE와 마찬가지로 코퍼스로서 갖춰야 할 이러한 대표성을 국내외

연구자들로부터 인정받았기 때문에 2012년 3월 31일 공개된 이후 영국(University of Birmingham 등)과 미국(Georgetown University, Northern Arizona University, Pennsylvania State University, University of Florida 등)을 포함한 국내 30개 이상 대학교, 100명 이상의 연구자들이 YELC 2011 기반으로 한 연구를 하고 있고 그 수는 매 년 늘고 있다.

3. YELC 2011의 실제성

Stubbs[28]는 학습자에 대해 언어학자나 영어 교사의 인위적인 간섭이 배제돼야 비로소 실제성(authenticity)을 가진다고 주장한다. YEPT가 연세대학교 내부적으로 개발됐다는 사실을 고려했을 때, 수험생은 이 시험에 대해 미리 준비하지 못한 상태이다. 또한, 문제은행 방식으로 출제되는 YEPT가 외부에 공개되지 않은 상태에 있다는 점을 고려하면, 수험생은 시험 당일 그 상태에서 본인들의 영어 능력을 그대로 보여줄 수밖에 없다. 시험 상황이기 때문에 언어학자나 영어 교사의 인위적인 간섭을 포함, 전자사전이나 인터넷의 도움을 받을 수 없다는 점도 수험생의 실제 영어를 보여주는 데 이바지한다. 결과적으로 YELC 2011을 구성하는 텍스트는 영어 학습자의 쓰기에 직·간접적으로 영향을 줄 수 있는 일체의 도움 없이 그 자리에서 본인의 영어 쓰기 능력을 그대로 보여준다는 점에서 실제성이 있다고 할 수 있다. 한편, 이와 같은 실제성을 바탕으로 생산된 텍스트는 제한된 시간 없이 집에서 작성한 쓰기 결과물과 달리 한국인 영어 학습자가 영어 쓰기를 할 때 범하는 대·소문자 사용의 오류, 철자법 오류, 구두법 사용의 오류 등도 고스란히 보여준다.

4. YELC 2011의 규모

코퍼스 연구자들 사이에서도 코퍼스를 구축할 때 어느 정도의 규모가 적당한지에 대한 논란은 지금까지도 끊이지 않는다[29]. Corpus of Contemporary American English(COCA)는 약 4억 2천 5백만 단어, Bank of English(BOE)는 5억 5천만 단어에 달할 정도이다. 영어 학습자 코퍼스는 영어 학습자가 직접 작성하거나 말한 내용을 기반으로 만들어진다. 따라서 텍스트가 다양한 영어 원어민 코퍼스보다 구축과정이 상대적으로 복잡하고, 텍스트를 구하기도 쉽지 않다. 따라서 앞서 언급한 LLC나 Cambridge Learner Corpus(CLC)의 경우 그 규모가 BNC, BOE, COCA의 10분의 1 수준인 1천만 단어와 2천3백만 단어 정도이다.

LLC나 CLC는 상업적 목적을 위해 세계적인 출판사의 거대 자본과 기술력이 투입된 경우이고, 개인연구자가 이런 규모의 영어 학습자 코퍼스를 구축하기란 현실적으로 거의 불가능하다. 이는 벨기에 루벤대학교에서 소개하는 세계 영어 학습자 코퍼스 리스트에서도 확인할 수 있는데, 다수의 학습자 코퍼스의 규모는 1백만 단어 이하로 나타났다. 국내에는 안성호와 이은영[30]이 구축한 한국인 영어 학습자 코퍼스가 13,384단어, Lee[31]의 학습자 코퍼스는 214,363단어이고, 앞서 언급한 SKELC의 경우 899,505단어 규모이다. 따라서 코퍼스 개발자에게 주어진 시간과 제약을 고려할 때, 영어 학습자 코퍼스는 학습자의 일반적인 오류를 연구할 수 있는 최소 1백만 단어 이상 규모이면 적절하다는 것이 저자들의 주장이다. 비근한 예로, 연구목적을 위해 사용할 수 있는 최대 규모의 개방형 영어 학습자 코퍼스로 알려진 스웨덴 University of Uppsala에서 구축한 The Uppsala Student English Corpus (USE)의 경우도 약 1백만 단어 정도이다[32].

IV. YELC 2011 구축과정

YELC 2011 구축에 사용된 실물 언어자료는 2011년 1월과 2월 사이에 시행된 YEPT 쓰기 시험에 참여한 총 3,563명의 응시자가 작성한 7,126개의 쓰기 결과물 중, 본 연구목적에 맞는 3,286명의 쓰기 결과 중 파트 2와 3, 총 6,572개의 텍스트이다. 이 장에서는 이 같은 코퍼스 구축에 대한 실제 과정을 단계별로 알아본다.

1. 설계 및 필터링

YELC 2011 구축에 대한 목적이 정해진 후, 본격적인 코퍼스 구축에 앞서 보유 텍스트 데이터베이스를 어떻게 코퍼스화 할지에 대한 논의가 이어졌다. YELC 2011 구축에 사용된 초기 자료는 마이크로소프트에서 개발

한 엑셀(Excel) 파일로 만들어진 데이터베이스 형태였다. 그 안에는 학생들의 기본 정보, 쓰기 결과물, 채점자, 채점결과 등이 각각의 셀에 독립적인 형태로 담겨 있다. 저자들은 이 엑셀 문서에 있는 학생들의 쓰기 결과를 코퍼스화하기 위해 각각의 쓰기 결과를 개별적인 텍스트 파일 형태로 만들었다. 이를 위해 마이크로소프트에서 개발한 노트패드(Notepad) 6.1 버전을 사용해서 엑셀 문서에 있는 3,563명의 쓰기 결과물 전체를 수작업으로 하나씩 개별 텍스트 파일로 만들었다.

그 결과 총 7,126개의 개별 텍스트 파일이 만들어졌고, 각 파일을 구별하기 위해 컴퓨터 하드드라이브에 편의상 YELC 2011_Part_01과 YELC 2011_Part_02라는 파일 폴더를 만든 후, 내부적으로 각 파일을 '수험번호_Writing Part'의 형식으로 이름 붙였다. 예를 들어 'AABB30235_01'인 파일은 AABB30235 학생이 작성한 쓰기 파트 1이다.

본 연구에서는 한국인 영어 사용자의 영어에 초점을 맞췄기 때문에 영어 이외의 언어가 있는 경우는 필터링을 통해 YELC 2011에 포함하지 않았다. 실제 학생들이 작성한 텍스트에는 영어 이외의 한국어, 불어 등이 포함된 텍스트가 다수 있었다. 주로 어려운 단어(e.g. 단백질), 표현(e.g. 자원하다), 고사성어(e.g. 백해무익) 등을 한국어를 영어로 못 바꾼 경우이다. 몇몇 학생들은 질문과 무관한 대답을 한 경우가 있었다. 이런 경우는 보통 두 가지 형태를 띠고 있었다. 첫 번째는 'I'm sorry I'm not very well English', 'I'm sorry I don't speak English'와 같이 질문과 무관하게 영어로 대답한 경우, 두 번째는 '죄송합니다. 열심히 배울게요', '신체검사에서 공익판정을 받았습니다.' 등과 같이 질문과 무관하게 한국어로 대합을 한 경우가 있었다. 두 경우 모두 질문에 대한 대답이 아닌 것으로 판단하고 YELC 2011에 포함하지 않았다.

전수검사를 하는 도중 다수의 학생이 이모티콘이나 불필요한 기호를 사용하고 있는 것도 발견됐다. 대표적인 예로는 ^^, :), ^^;, *.+ 등이 있다. 텍스트에서 나타나는 이와 같은 현상은 본 연구에서 추구하는 목적과는 거리가 있었기 때문에 YELC 2011에서 배제했다. 한편 YELC 2011은 최대 100단어 이내에서 대답하는 파트 1과 최대 300단어 이내에서 대답하는 파트 2가 한 세트를 이룬다. 따라서 두 개의 질문 중 한 개라도 대답하지 않은 경우는 세트를 이룰 수 없으므로 역시 YELC 2011 구축과정에서 제외했다.

2. 불순물 제거 및 무기명화

필터링하는 과정에서 저자들은 YEPT 시스템 개발자에 의해 기술적인 목적으로 텍스트에 자동으로 들어간 보이지 않는 컴퓨터 코드는 본 연구목적과 관련이 없어 '불순물'로 규정하고 전수 작업을 거쳐 하나씩 제거했다.

마지막으로, 수험생의 입학예정학과와 개인신상을 추측할만한 의미를 담고 있던 응시자 번호를 무기명화하는 작업을 진행했다. 저자들은 이를 위해 오픈 소스 파일 매니저 프로그램인 넥석스파일(NexusFile 5.3.1.5460)의 고급기능을 사용해서 최종 3,286명이 작성한 쓰기 결과물을 1번부터 3,286번까지 번호를 부여했고, 파트 1과 파트 2를 구분하기 위해 그 뒤에 파트 1인 경우 01, 파트 2인 경우 02를 붙였다. 예를 들어 1_01인 경우 학생 1번이 쓴 파트 1이고, 1_02는 학생 1번이 쓴 파트 2이다.

V. 결과

YELC 2011은 2011년 연세대학교 입학예정자 총 3,286명(남학생 1,958명, 여학생 1,328명)이 2011년 1월부터 2월 사이에 치른 영작문 시험에서 100자 이내로 작성해야 하는 영작문과 300자 이내로 작성해야 하는 영작문을 모아 구축한 한국인 영어 학습자 코퍼스이다. YELC 2011은 학생 번호, 성별, 등급에 대한 정보를 담고 있는 엑셀파일(YELC_2011.xlsx)과 총 6,572개의 텍스트 파일을 담고 있는 YELC_2011 파일 폴더로 구성되어 있고, 이 두 가지는 하나의 압축 파일 형태(YELC_2011.zip)로 연세대학교 영어코퍼스연구실을 통해 무료로 배포되고 있다.

YELC 2011을 요약하면, YELC 2011을 구성하는 텍스트는 품사나 저자 인적사항 등 부가정보를 담고 있지 않은 원시 코퍼스(raw corpus)이다. 또한, YELC 2011

은 한국인 영어 학습자들이 쓴 텍스트를 모았으므로 영어 학습자 코퍼스(English learner corpus)이다. YELC 2011을 구성하는 텍스트는 2011년 1월부터 2월이라는 한시적인 기간에 모았기 때문에 공시 코퍼스(synchronic corpus)로 볼 수 있고, 구축과정에서 영어 이외의 언어가 담겨있는 텍스트는 배제했기 때문에 단일어 코퍼스(monolingual corpus)이다.

YELC 2011 기여자의 연령대는 1995년생인 만 17세부터 1972년생인 만 40세까지 다양하다. 이들이 공부하게 될 곳은 문과대학, 상경대학, 경영대학, 이과대학, 공과대학, 음악대학, 의과대학, 치과대학, 간호대학 등 총 57개 학과, 학부 및 계열을 망라한다. YELC 2011 기여자 중 하위등급부터 상위등급 학생 수를 살펴보면 A1(최하등급)이 41명, A1+가 185명, A2가 684명, B1이 1,173명, B1+가 705명, B2가 378명, B2+가 81명, C1이 37명, C2(최고등급)가 2명이다. 상위등급인 C1과 C2가 상대적으로 적은 이유는 영어원어민이거나 영어실력이 뛰어나서 YEPT 자체를 면제받은 학생은 시험을 치르지 않았고, 중간에 필터링하는 과정에서 제외됐을 수도 있다.

VI. 분석 및 논의

1. YELC 2011의 표준화된 타입/토큰 비율

YELC 2011을 구성하는 텍스트 분석을 위해 WordSmith Tools 6.0[33](이하 '워드스미스')을 사용하였다. [표 1]은 워드스미스를 사용해서 YELC 2011을 구성하는 텍스트에 대한 통계적 정보를 보여준다. 이 표에서 볼 수 있듯이, YELC 2011의 파트 1에서 나타난 총 토큰(token) 수는 315,317개이고, 파트 2는 총 770,511개의 토큰으로 이루어져 있다. 이 중에서 타입(type)은 파트 1에서 11,308개, 파트 2에서는 16,416개가 있는 것으로 나타났다. YELC 2011의 토큰 값이 파트 1과 2를 더한 값과 같지만, YELC 2011의 타입 값이 파트 1과 2를 더한 값과 다른 이유는 파트 1과 2가 합쳐지면서 겹치는 유형은 타입에서 배제되기 때문이다.

일반적으로 텍스트에 사용된 어휘의 다양성을 살펴보기 위해 타입/토큰 비율(Type/Token Ratio, TTR) 값을 구한다[34]. 하지만 6,572개의 텍스트로 이뤄진 YELC 2011의 경우 텍스트의 길이가 일정하지 않기 때문에 일반적인 TTR 값은 다소 무의미할 수 있다. 따라서 텍스트의 길이를 인위적으로 일정하게 만들어 표준화된 타입/토큰 비율(Standardized Type/Token Ratio, STTR) 값으로 살펴볼 필요가 있다[35]. 이때 YELC 2011를 구성하는 전반적인 텍스트의 길이를 고려하고, 동시에 워드스미스의 STTR 베이스 최소 단위인 50을 기준으로 값을 구하면, 파트 1은 73.37%이고 파트 2는 76.79%이다. 따라서 전체 YELC 2011의 STTR은 75.93%로 구해진다.

표 1. YELC 2011의 통계 정보

	파트 1	파트 2	YELC 2011
텍스트	3,286	3,286	6,572
토큰	315,317	770,511	1,085,828
타입	11,308	16,416	21,839
표준화된 타입/토큰 비율	73.38	76.79	75.93
총 문장수	25,386	52,814	78,200
한 문장당 평균 단어 수	12.36	14.57	13.85

YELC 2011의 STTR을 비교하기 위해 연령대가 비슷한 영어 원어민 코퍼스 중 하나인 Louvain Corpus of Native English Essays(LOCNESS)를 사용하였다[36]. LOCNESS는 미국과 영국 대학생 영어 원어민들의 작문과 영국 대학교 입학 준비생인 영어 원어민 학생들의 작문을 모두 포함한다. 따라서 본 연구에서는 연령대가 비슷한 영국 대학교 입학 준비생인 영어 원어민 학생들의 작문만 추출해서 LOCNESS-E라는 서브코퍼스(sub-corpus)를 만들었고, LOCNESS-E를 YELC 2011과 비교했다. 워드스미스를 사용하여 LOCNESS-E를 YELC 2011과 같은 조건에서 STTR 값을 구했고, 그 값은 80.77%로 나타났다. 앞서 언급한 BAWE의 경우 비록 연령대는 다르지만, 영국 대학교 재학생들의 B+ 이상을 받은 상위 과제물을 모은 코퍼스이다. BAWE의 STTR 기준값을 YELC 2011과 같은 50으로 설정한 뒤

BAWE의 STTR을 구하면, 그 값은 78.40%가 나온다. 통계적으로만 봤을 때, YELC 2011 기여자의 STTR은 비슷한 연령대의 영어 원어민 학생들보다는 약 5%가량 낮게 나오지만, 영국 대학교에 재학 중이던 대학생들의 STTR과는 3%의 차이를 보이는 것으로 나타났다.

YELC 2011을 구성하는 총 문장 수는 78,2050이며 한 문장당 평균 13.85단어가 사용된 것으로 나타났다. 참고로 LOCNESS-E는 한 문장당 평균 12.36단어, BAWE는 한 문장당 평균 22.55단어가 사용된 것으로 나타났다. 통계적인 숫자만 보면, 한국 예비 대학생이 비슷한 연령대의 영어 원어민보다 평균적으로 한 문장을 완성하기 위해 더 많은 단어를 사용하고 있음을 알 수 있다. YELC 2011과 LOCNESS-E 모두 BAWE와는 상당한 차이를 보이는데, BAWE는 최소 500단어에서 많은 것은 10,000단어와 참고문헌까지 들어간 학술 보고서를 다수 포함한 과제물로 이루어져 있기 때문이라고 추정할 수 있다.

2. YELC 2011의 등급별 어휘 분포율 및 특성

YELC 2011을 구성하는 텍스트의 등급별 어휘 분포를 분석하기 위해 무료 소프트웨어 AntWordProfiler 1.4.0w을 사용하였다[37]. AntWordProfiler 1.4.0w에는 West[38]가 제안한 General Service List(GSL) 2,000 기본 어휘가 탑재되어 있는데, GSL 1K(1,000)에는 5백만 단어마다 332번 이상 나오는 고빈도 어휘가 1번부터 1,000번째까지 포함되어 있고, GSL 2K(2,000)에는 그 다음 고빈도 어휘가 순서대로 탑재되어있다[39]. AntWordProfiler 1.4.0w에는 Coxhead[40]가 제안한 학술어휘목록(Academic Word List, AWL) 570개의 단어족(word families)도 포함하고 있는데, AWL은 다양한 전공에서 사용되는 학술어휘가 표제어(headword)로 탑재되어 있다. 예를 들어 ABANDON이란 표제어는 ABANDONED, ABANDONING, ABANDONMENT, ABANDONS을 포함한다[41].

[표 2]는 YELC 2011의 등급별 어휘가 GSL 1K, GSL 2K, AWL, 이상의 어휘 목록에 없는 어휘를 포함하는 비율을 보여준다.

표 2. YELC 2011의 등급별 어휘 분포율

	A1	A1+	A2	B1	B1+	B2	B2+	C1	C2	ALL
GSL 1K	81.76	83.57	84.30	84.44	84.46	83.94	83.60	82.23	80.31	84.26
GSL 2K	6.18	5.77	5.89	5.85	5.96	6.36	6.50	6.56	8.76	5.98
AWL	2.33	2.56	3.00	3.14	3.29	3.74	3.76	4.44	5.76	3.25
Not in the list	9.73	8.10	6.81	6.56	6.29	5.97	6.14	6.78	5.16	6.51

[표 2]에서 볼 수 있듯이 학생들이 사용한 어휘는 대부분 GSL 1K+2K의 어휘목록에 약 90% 가까이 포함이 된 것으로 나타났다. 어휘목록별로 보면, 가장 낮은 등급인 A1의 GSL 1K는 81.76%이지만 중간 등급인 B1+ 등급까지는 계속 상승세를 보이다가 B2 등급부터는 점차 하락세를 보이며, 또 B2 등급부터는 GSL 2K의 어휘를 더 사용하는 추세를 보인다. 이것은 등급이 올라갈수록 학생들이 쉬운 어휘 사용이 줄고 그보다 조금 어려운 어휘를 사용하고 있다는 것을 확인할 수 있다. AWL의 경우는 낮은 등급에서 높은 등급으로 갈수록 학생들이 일관되게 AWL 어휘목록에 포함된 어휘를 많이 사용하는 것으로 드러났다. 또한, GSL 1K+2K, AWL에 포함되지 않은 'Not in the list(NITL)'가 낮은 등급에서 높은 등급으로 갈수록 일관성 있게 줄어드는 경향도 볼 수 있는데, NITL에서 줄어든 비율은 등급이 올라갈수록 GSL 2K와 AWL의 비율로 상쇄되는 경향을 보였다.

표 3. YELC 2011의 등급별 STTR

	A1	A1+	A2	B1	B1+	B2	B2+	C1	C2	Over all
STTR	75.31	74.94	76.23	77.45	78.29	79.23	79.84	80.31	80.88	77.70

[표 3]에서 볼 수 있듯이 워드스미스로 YELC 2011의 등급별 텍스트를 불러와 STTR 베이스를 50으로 설정을 한 후 값을 구하면, 등급이 높을수록 STTR이 높아지는 경향을 볼 수 있다. 최하등급인 A1(75.31%)과 최고등급인 C2(80.88%)와는 5% 정도의 차이가 있음을

확인할 수 있다. C2 등급을 제외하고는 앞서 언급한 비슷한 연령대의 영어 원어민 학생들(80.77%) 보다 낮은 수치를 보였다. 하지만 B2 등급부터는 영국 대학교 재학생들(78.40%) 보다는 높은 수치를 보였다. 한 문장당 평균 단어 수 역시 등급이 올라갈수록 늘어나고 있는 것도 확인할 수 있다. 특히 최고등급인 C2와 최하등급인 A1은 8단어 차이를 보이고 있는데, 이것은 학생들의 영어 쓰기 등급이 높아질수록 상위등급 학생들은 한 문장을 만들 때 더 많은 단어를 사용하고 있는 것을 보여준다.

VII. 결론 및 제언

본 연구는 연구 목적으로 사용될 수 있는 대규모 개방형 한국인 영어 학습자 코퍼스를 구축하고, 이를 통해 직관이 아닌 예비 대학생이 작성한 실제 작문을 기반으로 영어능력 등급이 높은 학생일수록 문장을 만들 때 같은 단어의 반복적 사용을 자제하고 상대적으로 어렵고 다양한 단어를 사용하고 있음을 확인할 수 있었다. 또한, 영어능력 등급이 높은 학생일수록 한 문장을 만들 때 평균적으로 더 많은 단어를 사용하고 있다는 점도 연구를 통해 발견할 수 있었다. 본 연구를 통해 우리나라 예비 대학생들의 어휘의 다양성 측면을 비슷한 연령대의 영어 원어민 학생들이나 영국 대학교에 재학중인 학생들과 비교했을 때 현저한 차이는 보이지 않는다는 것도 확인할 수 있었다.

YELC 2011의 활용은 다양하다고 할 수 있다. 우선 우리나라 고등학교 졸업자나 이와 동등한 자격을 갖춘 학생들이 자주 저지르는 오류분석을 연구할 수 있다. 예를 들어 '한국인들은 관사를 자주 틀린다'는 통설을 실제 대규모 한국인 영어 학습자 코퍼스인 YELC 2011을 통해 제한적으로나마 검증할 수 있다. 또한, YELC 2011을 이용해서 또래 영어 원어민 학생들보다 상대적으로 과도하게 사용하거나, 혹은 미흡하게 사용하는 표현들에 대한 비교 분석연구도 할 수 있다. 앞서 언급하였듯이 2012년 3월 31일 YELC 2011이 최초 공개된 이후, 국내외 연구자들은 YELC 2011을 기반으로 L1과 L2 writing 비교, 제2언어 습득 관련 조동사 사용, 영어 교육과정 어휘와 비교연구 등의 주제를 갖고 학술지, 콘퍼런스 발표지, 학위논문, 서적 출간을 목표로 다양한 연구 분야에 활용하고 있다.

참 고 문 헌

[1] J. Sinclair, *Corpus, concordance, collocation.* Oxford: Oxford University Press, 1991.

[2] A. O'Keefe, M. McCarthy, and R. Carter, *From corpus to classroom: Language use and language teaching,* Cambridge: Cambridge University Press, 2007.

[3] S. Hunston, *Corpora in applied linguistics,* Cambridge: Cambridge University Press, 2002.

[4] 양옥렬, 강창규, 남명우, "대화형 코퍼스의 설계 및 구조적 문서화에 관한 연구", 한국콘텐츠학회 논문지, 제4권, 제4호, pp.1-10, 2004.

[5] 하명정, "코퍼스에 기반한 문학텍스트 분석", 한국콘텐츠학회논문지, 제13권, 제9호, pp.447-447, 2013.

[6] 권혁승, 정채관, *코퍼스 언어학*, 한국문화사, 2012.

[7] H. Kucera and W. Francis, *Computational analysis of present-day American English,* Providence, R.I.: Brown University Press, 1967.

[8] P. Crawford, B. Brian, and H. Kevin. In H. Hamilton, W. Y. Chou (eds.), *The Routledge Handbook of Language and Health Communication,* Abingdon, UK: Routledge, pp.75-90, 2014.

[9] G. Kjellmer, *A dictionary of English collocations based on the Brown Corpus,* Oxford: Clarendon Press, 1994.

[10] G. Leech, "100 million words of English: the British National Corpus (BNC)," Language Research, Vol.28, No.1, pp.1-13, 1992.

[11] G. Leech, P. Rayson, and A. Wilson, *Word frequencies in written and spoken English: Based on the British National Corpus,* London:

Longman, 2001.

[12] P. Baker, A. Hardie, and T. McEnery, *A glossary of corpus linguistics,* Edinburgh: Edinburgh University Press, 2006.

[13] S. Granger, The computer learner corpus: A versatile new source of data for SLA research, In S. Granger (ed.), *Learner English on computer,* Abingdon, UK: Routledge, pp.3-18, 2013.

[14] C. James, Awareness, consciousness and language contrast. In. C. Mair, and M. Markus (eds.). *Proceedings of the new departures in contrastive linguistics conference*, Leopold - Franzens - University of Innsbruck, Austria, pp.183-197, 1992.

[15] S. Granger, "The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research," TESOL Quarterly, Vol.37, pp.538-546, 2003.

[16] P. Gillard and A. Gadsby, Using a learners' corpus in compiling ELT dictionaries, In S. Granger (ed.), *Learner English on Computer*, London: Longman, pp.159-171, 1998.

[17] 권혁승, "코퍼스 언어학의 실제 및 응용",응용언어학, 제24권, 제3호, pp.1-30, 2008.

[18] J. M. Choi, *Personal communication,* September 24, 2011.

[19] 한나래, 이수화, "학습자 코퍼스를 이용한 영어 전치사 오류 교정 모델 개발", 언어학, 제53권, 제1호, pp.163-185, 2009.

[20] N. R. Han, *Personal communication,* February 25, 2012.

[21] H. S. Kwon, "The SNU Korean learner corpus of English: Compilation and application," English Language and Linguistics, Vol.28, pp.203-228, 2009.

[22] H. K. Lee, "Investigating the applicability of the CEFR to a placement test for an English language program in Korea," English Language and Linguistics, Vol.17, pp.29-60, 2011.

[23] D. Biber, *University language: A corpus-based study of spoken and written registers,* Amsterdam: John Benjamins Publishing, 2006.

[24] T. McEnery, R. Xiao, and Y. Tono. *Corpus-based language study: An advanced resource book,* Abingdon, UK: Routledge, 2006.

[25] S. Alsop and H. Nesi, "Issues in the development of the British Academic Written English (BAWE) corpus," Corpora, Vol.4, pp.71-83, 2009.

[26] C. K. Jung and S. Wharton, "Finding textual examples of genres: Issues for corpus users," Korean Journal of English Language and Linguistics, Vol.12, No.1, pp.64-82, 2012.

[27] H. Nesi and S. Gardner, *Genres across the disciplines: Student writing in higher education,* Cambridge: Cambridge University Press, 2012.

[28] M. Stubbs, *Text and corpus analysis,* Oxford: Blackwell, 1996.

[29] N. Pravec, "Survey of learner corpora," ICAME Journal, Vol.26, pp.81-114, 2002.

[30] 안성호, 이은영, "한국인 학습자 전자우편 영어의 말뭉치 언어학적 분석", 영어학, 제5권, 제4호, pp.733-756, 2005.

[31] E. J. Lee, "Degree adverbial collocations in the Korean EFL learners' writing corpus: With a focus on intensifiers," Foreign Language Education, Vol.13, pp.1-21, 2006.

[32] M, Axelsson, "USE-The Uppsala Student English Corpus: An instrument for needs analysis," ICAME Journal, Vol.24, pp.155-157, 2000.

[33] M. Scott, *WordSmith Tools version 6*, Liverpool: Lexical Analysis Software, 2012.

[34] P. Scholfield, *Quantifying language: A researcher's and teacher's guide to gathering language data and reducing it to figures,*

Clevedon, Avon: Multilingual Matter, 1995.

[35] E. Castello, Integrating learner corpus data into the assessment of spoken interaction in English in an Italian university context, In S. Granger, G. Gilquin, and F. Meunier (eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead,* Louvain-la-Neuve: Presses universitaires de Louvain, pp.61-74, 2013

[36] S. T. Gries, and A. S. Adelman, "Subject realization in Japanese conversation by native and non-native speakers: Examplifying a new paradigm for learner corpus research," In J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2014: New Empirical and Theoretical Paradigms*, pp.35-54, 2014.

[37] L. Anthony, AntWordProfiler 1.4.0w Tokyo: Waseda University, 2013.

[38] M. West. *A general service list of English words,* London: Longman, 1953.

[39] I. S. P. Nation and L. Anthony, "Mid-frequency readers," The Journal of Extensive Reading, Vol.1, pp.5-16, 2013.

[40] A. Coxhead, "A new academic word list," TESOL Quarterly, Vol.34, pp.213-238, 2000.

[41] L. Bauer and I.S.P. Nation, "Word families," International Journal of Lexicography, Vol.6, No.3, pp.1-27, 1993.

저 자 소 개

이 석 재(Seok-Chae Rhee) 정회원

- 1988년 : 연세대학교 영어영문학과(문학사)
- 1990년 : 연세대학교 영어영문학과(영어학석사)
- 1998년 : 미국 University of Illinois at Urbana-Champaign (Ph.D. in Linguistics)
- 2000년 ~ 현재 : 연세대학교 영어영문학과 교수
- 2011년 ~ 현재 : 연세대학교 언어연구교육원장

<관심분야> 음성학 & 음운론, 언어이론, 영어교육, 코퍼스언어학

정 채 관(Chae Kwan Jung) 정회원

- 2000년 : 영국 University of Birmingham 생산공학·일본어 (공학사)
- 2002년 : 영국 University of Warwick 공학경영학(이학석사)
- 2011년 : 영국 University of Warwick 응용언어·영어교육학 (교육학박사)
- 2012년 ~ 현재 : 한국교육과정평가원 영어교육센터 부연구위원

<관심분야> : 코퍼스 구축, 코퍼스 활용 영어교육 및 평가, 자료주도 영어학습, 특수목적영어