

백아이 코퍼스의 영어 문장 구조 분석*

이유리**. 윤규철***

〈차 례〉

1. 서론
2. 연구 방법
 - 2.1. 연구 대상 및 추출 방법
3. 결과
 - 3.1. 코퍼스 발화 문장의 단어와 문장 개수
 - 3.2. 주절의 5형식 문형 유형 분포
 - 3.3. 복문 내부 종속절의 5형식 문형 유형 분포
4. 결론

【국문초록】

이 연구의 목적은 우리나라 공교육 혹은 사교육 영문법 교육

* 본 논문은 2023년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2023S1A5B5A17085179).

** 연구책임자 및 제1저자. 영남대학교 영어영문학과 박사졸업

*** 교신저자. 영남대학교 영어영문학과 교수

현장에서 잘 알려진 5형식 문장 구조가 영어 자연발화 음성 코퍼스인 벅아이 코퍼스에 어떠한 빈도로 존재하는지 알아보기 위한 것이다. 연령과 성별에 따라 네 집단으로 나누어진 코퍼스의 모든 문장을 문장 분류 지침에 따라 단문, 중문, 복문으로 구분하였고, 복문은 종속절의 개수에 따라 추가 분류하였다. 이후 주절과 종속절의 모든 문장을 5형식 구조로 분류하였다. 분류 결과, 대다수의 문장이 1, 2, 3형식으로 나타났고 4, 5형식은 그 수가 매우 적었다. 1, 2, 3형식 문장 중에서 1, 2형식 문장은 비슷한 빈도를 보였고, 제일 빈도가 높은 구조인 3형식은 이들보다 거의 두 배에 가까운 빈도를 보였다. 이러한 결과는 영어 문장 구조 교육에 있어서 고려할 기초 자료로서 의의가 있다고 판단된다.

주제어: 벅아이 코퍼스, 문장, 발화, 문장 구조, 문형 분석, 영어

1. 서론

우리나라의 영문법 교육이 시작된 이래로 영문법 교육의 주류를 차지해 온 것은 어니언스의 5형식 문형 이론이었다.¹⁾²⁾ 이 이론은 영어 교육 과정이 여러 차례 개정되어 오면서 실용적이고 의사소통 능력 중심의 영어 교육을 추구하는 현재에 이르러서도 여전히 초등 중등 및 고등학교 대부분의 교과서와 참고서 및 시중 서점에

1) Onions, C. T., *An Advanced English Syntax*, London: Keagun Paul, 1932, pp. 4-38.

2) Onions, C. T., *Modern English Syntax* (New ed. Of *An Advanced English Syntax*, by B.D.H. Miller), London: Routledge & Keagun Paul, 1971.

있는 모든 영문법 교재의 중심 내용을 차지하고 있다. 영문법을 처음 배우는 거의 모든 학습자들은 교재의 초반부에 등장하는 5형식 문형 혹은 5형식 문장 구조를 빠짐없이 배워야만 한다.

이 이론은 문장의 표면 구조만을 중점적으로 보기 때문에 문장이 지니고 있는 내적 의미를 파악한다거나 분석할 때에 다소 문제를 야기시킬 수 있는 단점을 지니고 있음에도 불구하고³⁾⁴⁾, 영어를 외국어로서 학습할 때에 많은 교육적 장점을 지니고 있기 때문에 여전히 한국의 교육 실정에서 영문법 교육에 큰 역할을 수행하고 있음을 부정할 수 없다. 그러나, 영어에서 쓰일 수 있는 평서문의 모든 문장 형식을 다섯 가지로 간략하게 배울 수 있다는 장점이 있는 반면에, 1형식, 2형식, 3형식, 4형식, 5형식 문형이 각각 어느 정도의 빈도로 어떠한 경우와 상황에서 쓰이는지에 대한 정보는 문법 책에 제시되어 있지 않다. 이에 대한 통계적 조사를 실시한 연구가 존재하지 않기 때문일 것이다.

영어 문법 교재나 참고서에서는 주로 시험 대비를 위하여 우리말 문장 구조와는 상이한 정도가 크다고 볼 수 있는 4형식(3형식으로의 전환 등)이나 5형식(목적어와 보어의 관계 등)을 비중 있게 다루는 경우가 많다. 그러나 영어 학습의 목표가 영어로 자신의 생각을 자유롭게 의사소통 하는 것이라면, 다섯 가지 문형 모두를 배우는 것도 좋겠지만 어느 문형(들)이 상대적으로 훨씬 많이 사용되는지에 대한 통계적 빈도 조사에 근거하여 이들에 대한 교육과 학습에 더욱 노력을 들이는 것이 영어 의사소통 능력을 키우는 데에 보다 효율적이고 능률적이지 않을까 하는 고민도 필요하다고 생각된다.

3) 배영남, 「5형식 문형 이론과 영어 문법 교육」, 『언어과학연구』 24, 2003, pp. 83-110.

4) 정덕교, 「영어 문장형식의 연구 및 적용 - 영문법 5형식의 재조명」, 『교양교육연구』 4(2), 2010, pp. 177-204.

영어 교육과 관련한 코퍼스 기반의 연구들이 많이 진행되고 있지만, 대부분의 경우, 텍스트 코퍼스의 활용 유무에 연구의 초점이 집중되어 있다. 예를 들면, 문법 규칙에 대한 이해도를 연구한 임수영·이은주(2012)⁵⁾의 고등학생 대상 연구에서는 코퍼스 자료를 활용한 집단과 그렇지 않은 집단의 문법 이해도 정도를 비교하였고, 초등학생을 대상으로 한 김현주⁶⁾의 연구에서는 전통적 교육 방법과 교육용 코퍼스를 활용한 교육 방법과의 테스트 점수 및 참여도와 흥미도를 살펴보았으며, 코퍼스 활용 여부를 대학생들에 적용한 연구⁷⁾에서는 단순 문법보다는 복잡한 문법 사항 교육에 코퍼스가 효과를 나타내었다고 한다. 또한 영어 교사들이 코퍼스를 활용하면 쓰기지도에 도움을 받을 수 있다는 연구보고서⁸⁾도 있다.

코퍼스 내의 내용을 연구 대상으로 삼은 최근의 많은 연구들의 경우에도 영어 구동사에 대한 코퍼스 기반 연구⁹⁾나 영어 관계대명사 사용에 대한 코퍼스 기반 연구¹⁰⁾, 음성 코퍼스 내 명사구나 전치사구 등의 어휘 묶음 분석 연구¹¹⁾, 코퍼스 기반의 영어 주어-

5) 임수영·이은주, 「코퍼스를 활용한 문법 과제가 고등학생의 영어 문법 학습과 정의적 반응에 미치는 영향」, 『영어학』 12(2), 2012, pp. 303-325.

6) 김현주, 「용어색인을 활용한 영어 문법 학습의 효과와 영어 학습에 대한 태도 변화 분석」, 한양대학교 석사학위논문, 2016.

7) 정송휘, 「코퍼스를 활용한 문법 학습이 대학생의 문법 능력에 미치는 영향」, 한남대학교 박사학위논문, 2014.

8) 이문복·신동광·전유아·원장호·이희중·강동길, 「코퍼스 활용을 통한 영어 교사의 실제적 영어 쓰기지도 능력 향상 방안 연구」, 연구보고서, 한국교육과정평가원, 2008.

9) 김진란, 「고등학교 영어 교과서에 나타난 영어 구동사의 코퍼스 기반 분석」, 전남대학교 석사학위논문, 2019.

10) 배소영, 「학술적 에세이 코퍼스를 기반으로 한 한국 대학생과 영어 원어민 대학생의 관계대명사 which와 that의 사용 비교」, 이화여자대학교 석사학위논문, 2018.

11) 안효빈, 「한국 영어 학습자 구어 코퍼스 기반 어휘 묶음 사용 양상 분석 연

동사 도치 구문 연구¹²⁾, 코퍼스 내의 영어 동의 형용사 연구¹³⁾, 코퍼스에 나타난 담화표지어 like 연구¹⁴⁾ 등 문장 내부의 특정 항목들에 대한 연구가 주류를 이루고 있고 문장 자체의 구조나 문형에 대한 조사는 거의 없는 것으로 파악된다.

많은 연구 분야에서 빅데이터 구축에 심혈을 기울이고 있는 현 상황에서 언어 분야에서도 많은 텍스트 및 음성 코퍼스들이 구축되고 있는데, 영어 원어민들이 생산한 텍스트 및 음성 코퍼스에서 영어 문장의 기본적 문형과 구조에 대한 통계적 빈도나 분포 조사가 부족한 현 상황은 개선될 필요가 있다고 생각된다. 따라서 본 연구에서는 영어 구어체 음성 코퍼스인 백아이 코퍼스¹⁵⁾에 나타난 남녀노소 영어 원어민들의 발화 문장들에 대하여 5형식 문형 및 단문, 중문, 복문의 관점에서 문장 구조 유형의 분포와 빈도 조사를 통하여 영어 문법 교육에 있어서 어떠한 문장 유형에 보다 많은 자원과 노력을 기울여야 할 지에 대한 기초 조사를 수행하고자 한다.

구], 인천대학교 석사학위논문, 2022.

12) 서아리, 「영어 주어-동사 도치 구문에 관한 코퍼스 기반 연구」, 계명대학교 박사학위논문, 2022.

13) 한인석, 「코퍼스를 이용한 영어 동의 형용사 연구: “able, capable, competent & qualified”를 중심으로」, 『언어연구』 38(3), 2022, pp. 307-322.

14) 최인지, 「영국 영어 대화 코퍼스에 나타난 담화표지어 like 연구」, 『새한영어영문학』 64(4), 2022, pp. 149-174.

15) Pitt, M.A., Dille, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E, Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor), 2007.

2. 연구 방법

2.1. 연구 대상 및 추출 방법

영어 벅아이 코퍼스는 구어체의 자연 발화 음성 코퍼스로서 미국 오하이오 주에 거주하는 미국 중류층 남녀노소 40명을 각각 1시간씩 인터뷰하여 구축한 것이다. 약 30만 개의 단어가 변이음별로 레이블링이 되어 있으며 문장 텍스트도 함께 배포되지만, 음성 코퍼스에 시간대별로 마킹되어 있지는 않다. 음성을 듣지 않고 주어진 텍스트만 분석할 수도 있겠지만, 구어체의 특성상 발화 중간에 끊어지는 경우가 많고 불완전한 문장이 상당히 많아서 억양이나 휴지기 등의 운율 정보 없이 텍스트만으로 문장의 5형식 구조나 단문, 중문, 복문의 구조를 판단하기는 어려운 경우가 많아서, 40명 전체 40여 시간에 해당하는 255쌍의 사운드/텍스트그리드를 직접 들으면서 문장 구조를 파악하는 방법을 택하기로 하였다.

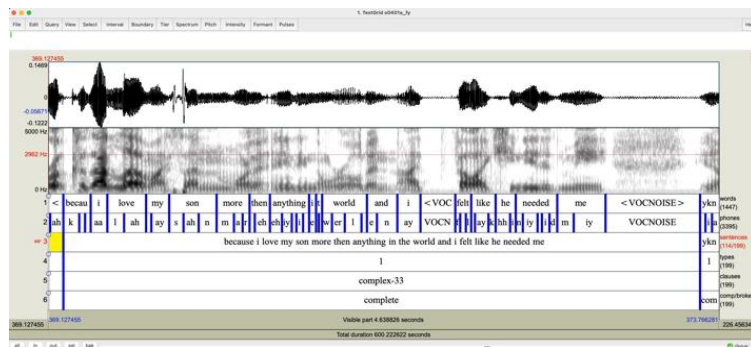
발화된 문장을 일관되고 신속하게 추출하기 위하여 프랏 소프트웨어¹⁶⁾를 활용하였고, 231줄 길이의 스크립트를 작성하여 코퍼스 처리 작업을 하였다. 이 스크립트는 벅아이 코퍼스에서 한 쌍의 사운드/텍스트그리드를 자동으로 탑재하여 사용자에게 보여주고 <그림 1>과 같은 포즈창을 띄워준다. 그 다음 사용자가 하나의 발화 문장이라고 판단하는 부분을 마우스로 선택하여 포즈창에 주절의 5형식 문형 유형과 종속절 혹은 대등절의 5형식 유형과 개수 등 필요한 정보를 입력하여 확인 버튼을 누르면 해당 정보를 텍스트그리드에 입력한 후 다음 문장을 선택하도록 해준다.

16) Boersma, Paul, "Praat, a system for doing phonetics by computer", *Glot International* 5:9/10, 2001, pp. 341-345.

〈그림 1〉 스크립트 실행 후 나타나는 포즈창

계속해서 사용자가 직접 음성을 들어보면서 발화 문장의 경계를 설정하고, 주절의 문형의 종류는 1~5형식(sentenceType) 중에서 선택하여 클릭하고, 종속절이 존재하는 복문인 경우 typeComplex 필드의 빈칸에 5형식 문형 종류를 절의 갯수만큼 판단하여 입력하고, 문장이 대등절로 구성된 중문인 경우typeCompound 필드에 5형식 문형 종류를 역시 절의 갯수만큼 입력하도록 하였다. 마지막 completeType 항목은 완전한 문장인 경우 complete, 문장이 끊기거나 중지된 경우 broken 두 항목 중 선택하도록 하였다. 이렇게 주절과 종속절의 5형식 문형과 대등 혹은 종속절의 문형 판단이 끝나면 텍스트그리드에 선택 구간의 모든 단어들이 문장으로 완성되어 문장층에 입력되고, 나머지 정보들은 해당 층에 스크립트

가 자동 입력하여 <그림 2>처럼 보이게 된다. 동일한 정보가 텍스트그리드별로 텍스트 파일로도 출력되어 추후에 엑셀 파일로 통합된다.



<그림 2> 문장 정보가 모두 입력된 텍스트그리드

스크립트로 출력된 255개의 텍스트 파일들은 엑셀 파일로 통합되어 모든 개별 문장에 대하여, 파일이름, 화자번호, 성별, 나이, 인터벌번호, 문장의 시작 시간, 끝 시간, 주절유형, 대등절유형, 종속절유형, 완전/불완전문장, 문장 내 단어 수, 문장텍스트, 단문/중문/복문유형 등의 정보가 기록되어 분석 절차에 들어간다.

이러한 추출 절차에 앞서 구어체의 특성상 발화 문장의 경계를 설정하기 위하여 ‘문장’에 대한 새로운 정의가 필요했다. 다음은 벅아이 코퍼스에서 발화 문장을 구분하기 위해 본 연구에서 설정한 지침들이다.

1. 한 단어 문장 “Yes, No, Right, Probably, Exactly” 등 제외.
2. 군말 “I mean, You know, You see, I guess, It’s like, I suppose, like, The thing is” 제외.
3. Let’s의 경우 5형식이 아닌 직후의 동사부터 구조 파악.

4. be able to, be willing to, be going to, would like to는 조동사 취급, 직후의 동사부터 구조 파악.
5. 대동사나 대부정사는 1형식으로 본다.
6. I was like, “did you...?”나 I go, “did you...?”처럼 be like이나 go가 say의 용법으로 쓰이면 3형식으로 본다.
7. 대등접속사 and, but, or 등이 문두나 문미에 있어도 전후에 휴지기가 있으면 대등절이 아닌 별도의 발화 문장으로 본다. 휴지기가 없으면 대등절로 본다.
8. 종속접속사 when, because, while, as 등이 전후의 절과 휴지기가 없으면 이어지면 복문으로 본다. 긴 휴지기가 있으면 별도 문장으로 본다.

이러한 지침들 없이 자연 발화 구어체 코퍼스의 문장을 선별하려 하면 여러 문제에 부딪히게 된다. 7번 대등절 판단의 경우, 구어체의 특성상 대등접속사가 매우 자주 씌없이 나오는 경우가 많은데 이를 모두 대등절로 보게 되면 엄청나게 길고 부자연스러운 대등절의 증문이 나타나게 된다. 해당 부분을 만약 문어체의 에세이로 썼다면 절대 하나의 증문으로 쓰지 않았을 것이기 때문이다. 주관적 판단을 가능한 없애고 객관적으로 판단하기 위해서 문장 전후의 휴지기 존재 유무가 매우 중요하게 된다. 휴지기가 없는 진정한 대등절을 가능한 많이 찾아내면서, 구어체의 특성상 대등절처럼 보이지만 실제로는 분리된 문장들을 구분하기 위해 7번 기준은 필요하다고 생각된다. 복문을 찾아낼 때에 필요한 8번 기준에도 마찬가지로 휴지기의 유무가 객관적 판단의 근거가 된다.

군말(discourse markers)이나 유사 조동사 같은 구에 가려진 5형식 문형을 가능한 많이 찾아내기 위해 2, 3, 4번의 지침도 필요하다. I mean이나 You know와 같은 군말을 문장의 일부로 인정하면 모든 문형이 3형식이 되어버려 그 이후에

나오는 구나 절의 진정한 문형이 종속절에 간혀 버리는 결과를 갖게 될 것이다. 또한 be able to, be willing to, be going to 같은 유사 조동사구도 문장의 일부로 인정하게 되면 모두 2형식 문형이 되어버려 그 후에 구나 절의 실제 문형이 가려지게 될 것이다. 따라서 이들을 문장 구조에 제외시킴으로써 그 후에 나오는 동사구 이하의 실제 문형을 드러내어 분석 대상에 포함시키고자 하였다. 그런 의미에서 3번 항목 let's의 경우도 5형식이 아닌 그 이후에 나오는 동사구의 구조를 드러내 문형을 판단하게 된다.

직접 인용 표현을 언급한 6번 항목은 be like(2형식), go(1형식)의 실제 문형이 아니라 say로 쓰인 3형식으로 판단했으며, 5번 대동사나 대부정사의 경우 보이지 않는 부분이 아니라 보이는 부분을 중심으로 판단한다는 지침이다. He had lunch like I did의 경우 did가 had lunch(3형식)를 대신하고 있지만 실제 보이는 것은 did라는 대동사이고 이는 표면상으로 목적어가 없으므로 1형식으로 판단하였다. 보이지 않고 실제 말하지 않은 것에 대하여는 문형으로 인정하지 않았고 보이는 것을 중심으로 문형을 판단한 것이다.

3. 결과

3.1. 코퍼스 발화 문장의 단어와 문장 개수

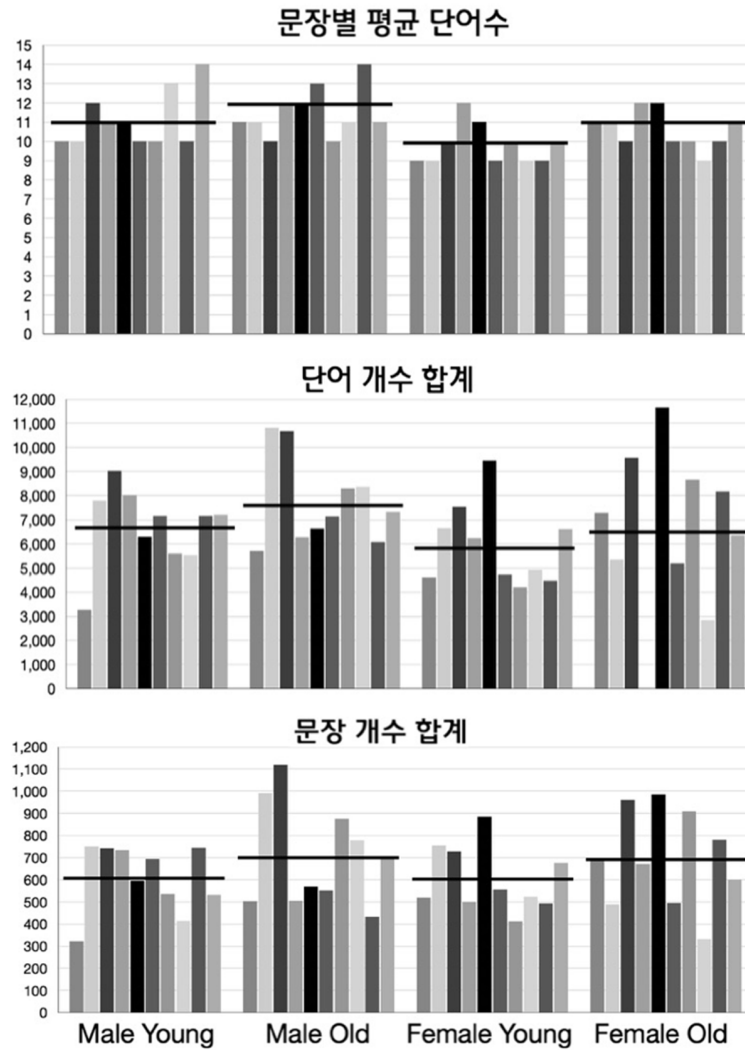
발화 문장 추출 지침을 바탕으로 코퍼스로부터 뽑아낸 영어 문장의 개수를 성별과 연령별로 정리하면 <표 1>과 <그림 3>과 같다. 총 문장 수는 26,051개이며, 총 단어 수는 277,153개이다.

〈표 2〉 코퍼스의 남녀노소별 화자별 단어, 문장 개수 통계

성별	연령	화자	문장별 단어수 (평균)	단어수 총합	문장수 총합
males	young	#06	10	3,258	323
		#11	10	7,800	750
		#13	12	9,028	742
		#15	11	8,027	735
		#28	11	6,294	596
		#30	10	7,166	693
		#32	10	5,592	535
		#33	13	5,527	414
		#34	10	7,151	745
		#40	14	7,201	532
		평균	11	6,704	607
	old	#03	11	5,705	504
		#10	11	10,823	991
		#19	10	10,670	1,120
		#22	12	6,282	506
		#23	12	6,633	569
		#24	13	7,133	552
		#29	10	8,313	875
		#35	11	8,364	779
		#36	14	6,075	433
		#38	11	7,326	694
		평균	12	7,732	702
females	young	#01	9	4,603	520
		#04	9	6,650	754
		#08	10	7,549	728
		#09	12	6,228	499
		#12	11	9,448	885
		#21	9	4,723	555
		#26	10	4,210	412
		#31	9	4,922	523
		#37	9	4,474	494
		#39	10	6,615	675
		평균	10	5,942	605

	old	#02	11	7,285	692
		#05	11	5,347	490
		#07	10	9,569	961
		#14	12	8,286	672
		#16	12	11,661	985
		#17	10	5,182	496
		#18	10	8,664	909
		#20	9	2,843	333
		#25	10	8,172	780
		#27	11	6,354	600
		평균	11	6,509	692

〈그림 3〉에서 보듯이 개인별 차이가 매우 심하게 나타나고 있는데, 개인의 음성 파일을 들어보면 사람에 따라 말수가 많은 사람과 적은 사람의 차이가 심한 편이며, 인터뷰어가 화자의 대답을 유도하는 기법에 따른 차이도 보이는 듯하다. 어떤 인터뷰어는 다른 인터뷰어보다 자신의 말이 상대적으로 많았고 화자의 대답을 유도하는 기교가 좀 부족해 보이는 경우도 있었다. 하지만, 이러한 인터뷰어의 성향을 고려하더라도 화자 개개인의 말수가 많고 적음이 전체 문장의 수에 큰 영향을 미치는 것은 분명해 보였다. 평균적으로는 대체로 한 문장 당 평균 11 단어 정도를 쓰고 있으며, 1 시간 대화 중에 평균 6천여 단어로 6백여 문장을 말하는 것으로 나타났다.



〈그림 3〉 코퍼스의 집단별 단어, 문장 통계. 수평바는 평균

3.2. 주절의 5형식 문형 유형 분포

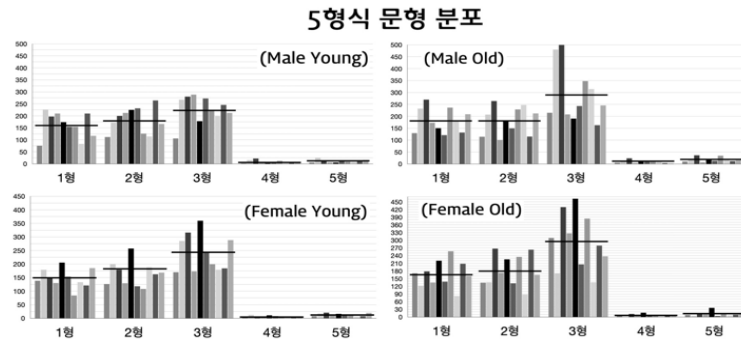
화자별로 주절의 문장 구조 분포를 5형식 문형별로 나타내 보면 <표 2>와 <그림 4>와 같다. 발화 도중에 문장이 중간에 끊긴 경우는 문형 파악 분석에서 제외하였고 종속절이 없는 단문도 주절의 분석에 포함시켰다.

<표 2> 코퍼스의 남녀노소 화자별 단어, 문장 개수 통계

성별	연령	화자	5형식 문형				
			1형	2형	3형	4형	5형
males	young	#06	76	112	106	2	5
		#11	226	172	267	15	25
		#13	197	200	280	22	15
		#15	210	212	288	4	5
		#28	174	225	178	7	5
		#30	154	231	272	6	11
		#32	154	126	222	12	9
		#33	83	114	199	2	8
		#34	210	264	246	5	11
		#40	117	166	212	3	16
		평균	160	182	227	8	11
	old	#03	129	114	215	5	10
		#10	232	207	480	13	17
		#19	270	264	501	24	36
		#22	173	101	208	4	7
		#23	150	184	190	10	16
		#24	121	149	243	10	13
		#29	236	229	348	12	35
		#35	181	247	314	11	16
		#36	132	115	163	4	12
		#38	209	212	246	2	15
		평균	183	182	291	10	18
		#03	129	114	215	5	10
		#10	232	207	480	13	17
		#19	270	264	501	24	36

	#22	173	101	208	4	7
	#23	150	184	190	10	16
	#24	121	149	243	10	13
	#29	236	229	348	12	35
	#35	181	247	314	11	16
	#36	132	115	163	4	12
	#38	209	212	246	2	15
	평균	183	182	291	10	18
	#03	129	114	215	5	10
	#10	232	207	480	13	17
	#19	270	264	501	24	36
	#22	173	101	208	4	7
	#23	150	184	190	10	16
	#24	121	149	243	10	13
	#29	236	229	348	12	35
	#35	181	247	314	11	16
	#36	132	115	163	4	12
	#38	209	212	246	2	15
	평균	183	182	291	10	18

〈표 2〉의 수치와 〈그림 4〉의 분포 유형에서 보듯이 주절의 문장 구조는 1~5형식 문형 중에서 거의 모든 화자들이 1, 2, 3형식 문형들을 주로 사용하였으며 4, 5형식 문형은 상대적으로 사용 빈도가 매우 낮음을 알 수 있다. 우리가 영문법 시간에 학습하는 다섯 개의 문형의 분포가 원어민의 구어체 사용 빈도에 있어서 서로 유사하거나 완만한 증감이 아닌 1~3형식의 집단과 4~5형식의 두 집단으로 비교해 보면 급격할 정도로 빈도 차이가 있다는 사실은 매우 흥미롭고도 의외의 발견이라고 생각된다. 4, 5형식이 이렇게 빈도가 낮다면 영어 문형 학습과 교육에 있어서 이러한 조사 결과가 어떤 의미가 있는지 교육적 측면에서 고려해 보아야 할 것으로 생각된다. 1~3형식 문형 집단에서는 대체로 1, 2, 3형식 문형의 순서로 개수가 증가하였고 3형식은 특히 1, 2형식 문형보다 그 수가 컸다.



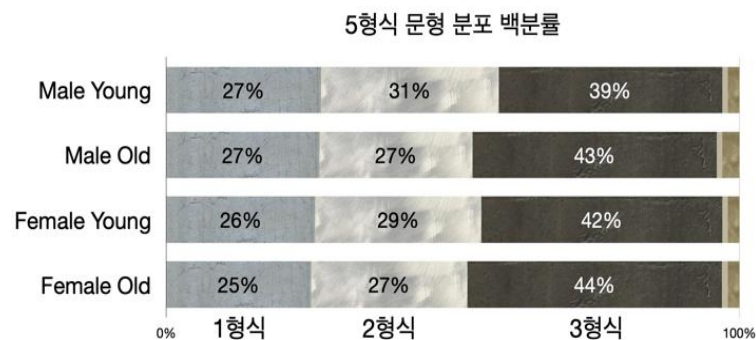
〈그림 4〉 코퍼스의 집단별 문형 분포. 수평바는 평균

주절의 5형식 문형 분포에 대한 통계 분석은 〈그림 4〉에서 보듯이 데이터 분포가 정규분포를 이루고 있지 않아 비모수 통계 중 집단간 비교 분석을 위한 방법인 Mann-Whitney U 검정을 R-4.4.4(R Core Team, 2020)을 기반으로 한 Rstudio(R Studio Team, 2019)에서 시행하였다. 우선 젊은(Young) 집단 내에서의 성별 요인의 영향을 분석한 결과 $W=2,310,135$, $p<0.05$ 로 5형식 문형 분포에 유의미한 영향을 미치고 있는 것으로 나타났다. 〈그림 4〉의 좌측 위아래 막대그래프가 이에 해당되는데, 1형식에서 3형식으로의 빈도 증가 폭에 있어서 다소 차이를 보이는 것으로 보인다. 나이 든(Old) 집단에 있어서도 5형식 문형 분포에 성별 요인이 유의미한 영향을 미치는 것으로 나타났다($W=3,547,278$, $p<0.05$).

이번에는 연령 요인이 미치는 영향을 알아보기 위해 남성(Male) 집단에 대하여 검정을 시행한 결과 $W=3,032,705$, $p<0.05$ 로 유의미한 영향을 미치는 것으로 나타났다. 〈그림 4〉의 위 좌우 두 막대그래프가 해당되는데, 3형식의 빈도가 상대적으로 높아 보인다. 여성(Female) 집단에 대한 분석에서는 $W=2,594,934$, $p>0.05$ 로 연령은 5형식 문형 분포에 영향을 미치지 않는 것으로 나타났다. 〈그림 4〉에서 아래 좌우 막대그래프를 보면 절대 빈도의

차이는 있지만 문형 분포의 패턴은 매우 유사해 보이는 것을 알 수 있다.

이 세 문형의 남녀노소 집단별 평균 백분율을 <그림 5>에 나타내었는데, 젊은 남성 집단을 제외하고는 1, 2형식은 대체로 20 퍼센트대에 존재하는 반면에 3형식은 40퍼센트 대에 존재하여, 빅아이 코퍼스의 영어 구어체에서 5형식 문형 중에서 3형식 문형 구조가 40%대의 압도적인 다수를 차지하고 있음을 알 수 있었다. 그 뒤로 2형식이 20% 후반대를 차지하고, 1형식은 20% 중후반대를 차지하였다. 4, 5형식은 1~3% 정도로 극소수를 차지하고 있음을 알 수 있었다.



<그림 5> 주절의 집단별 1, 2, 3형식 문형 백분율 (평균)

3.3. 복문 내부 종속절의 5형식 문형 유형 분포

문장의 구조를 단문, 대등절이 포함된 중문, 그리고 종속절이 포함된 복문의 세 집단으로 나누어 그 분포를 나타내 보면 <표 3>과 같다. 서두의 문장 분석 지침 7번에서 언급했듯이, 구어체의 특성상 문두나 문미에 대등접속사가 있어도 휴지기가 있으면 별도의 발화문장으로 보았기 때문에 휴지기가 없는 순수한 중문의 빈도는

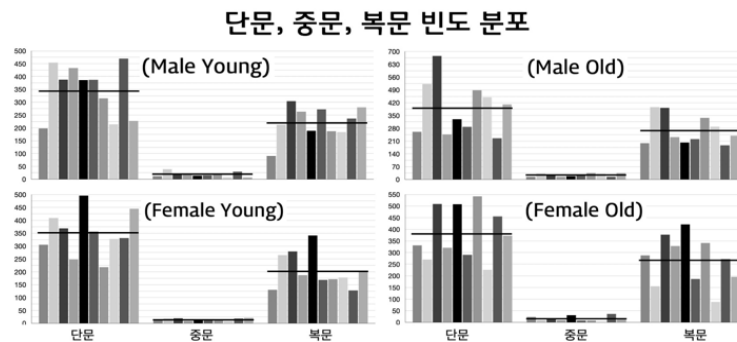
매우 적었다. 또한 주절의 5형식 문형 분석에서 중문 접속사의 앞에 있는 문장이 분석되었기 때문에 접속사 뒤의 문장들에 대한 분석은 따로 하지 않기로 한다. 대신 여기서는 단문과 복문의 빈도에 초점을 맞추어 구어체의 특성을 살펴보는 것이 유익하다고 본다.

〈표 3〉 코퍼스의 남녀노소 화자별 단문, 중문, 복문 통계

성별	연령	화자	단문	중문	복문
males	young	#06	198	12	91
		#11	454	39	212
		#13	388	22	304
		#15	433	23	263
		#28	386	14	189
		#30	387	15	272
		#32	315	21	187
		#33	214	8	184
		#34	470	30	236
		#40	227	7	280
		평균	347	19	222
	old	#03	261	14	198
		#10	522	32	395
		#19	676	27	392
		#22	247	15	231
		#23	330	18	202
		#24	288	27	221
		#29	488	35	337
		#35	449	31	289
		#36	225	14	187
		#38	411	34	239
		평균	390	25	269
females	young	#01	305	11	130
		#04	409	19	265
		#08	369	20	279
		#09	248	9	187
		#12	496	12	341
		#21	356	10	169
		#26	218	13	171

		#31	328	8	178
		#37	331	19	128
		#39	445	21	201
		평균	351	14	205
	old	#02	331	23	288
		#05	270	9	155
		#07	509	15	377
		#14	321	9	328
		#16	508	31	421
		#17	290	8	186
		#18	542	8	341
		#20	226	1	88
		#25	455	36	272
		#27	372	18	196
		평균	382	16	265

〈표 3〉과 〈그림 6〉에서 보듯이, 순수한 중문을 제외하고 단문과 복문을 비교해 살펴보면, 남녀노소 집단별로 개인차는 심하지만 전체적으로 단문의 비중이 상대적으로 큰데, 두 문형 사이의 비율은 대략 6:4 정도로 복문의 비중도 적지 않음을 알 수 있다.



〈그림 6〉 단문, 중문, 복문 빈도 분포. 수평바는 평균

젊은(Young) 집단에서 성별이 단문, 중문, 복문의 분포에 미치는 영향을 Mann-Whitney U 검정으로 알아본 결과, $W=17,780,650$, $p<0.05$ 로 유의미한 것으로 나타났지만, 나이 든 집단에서는 $W=24,150,359$, $p>0.05$ 로 성별이 영향을 미치지 않는 것으로 나타났다. 남성 집단에서 연령이 미치는 영향은 $W=21,685,120$, $p<0.05$ 로 유의미했으며, 여성 집단에서도 연령이 영향을 미치는 것으로 나타났다($W=21,778,119$, $p<0.05$).

단문의 5형식 문형 분포는 앞서 주절의 5형식 문형 분석에서 살펴보았기 때문에 이번에는 복문의 종속절 내부에서 5형식 문형의 분포는 어떠한지 살펴보도록 하자. 복문의 경우 종속절의 개수가 한 개 이상의 경우도 많다. 종속절이 하나인 경우부터 다섯 개 이상인 경우까지 종속절(들)의 5형식 문형 분포를 집단별 평균값으로 <표 4>와 <그림 7>에 나타내었다.

<표 4> 종속절의 개수에 따른 집단별 5형식 문형 분포

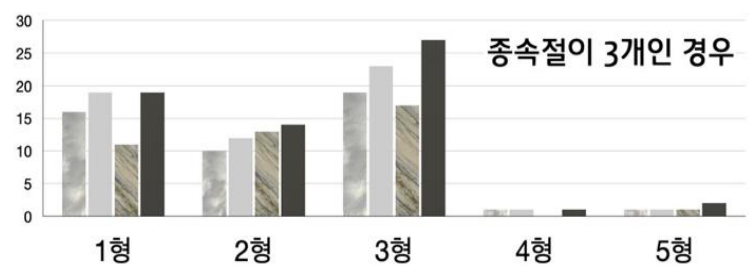
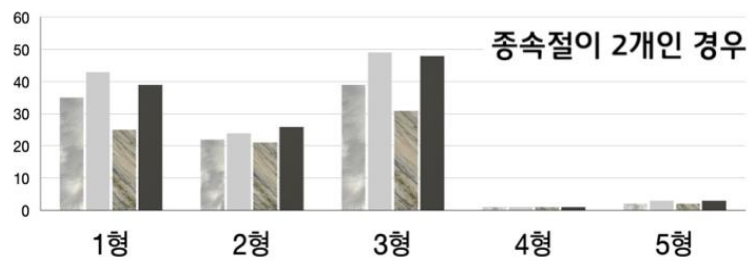
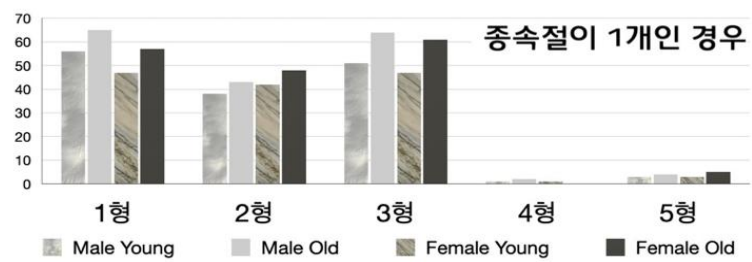
집단	절1개	5형식 문형				
		1형	2형	3형	4형	5형
males young	149	56	38	51	1	3
males old	177	65	43	64	2	4
females young	141	47	42	47	1	3
females old	172	57	48	61	0	5

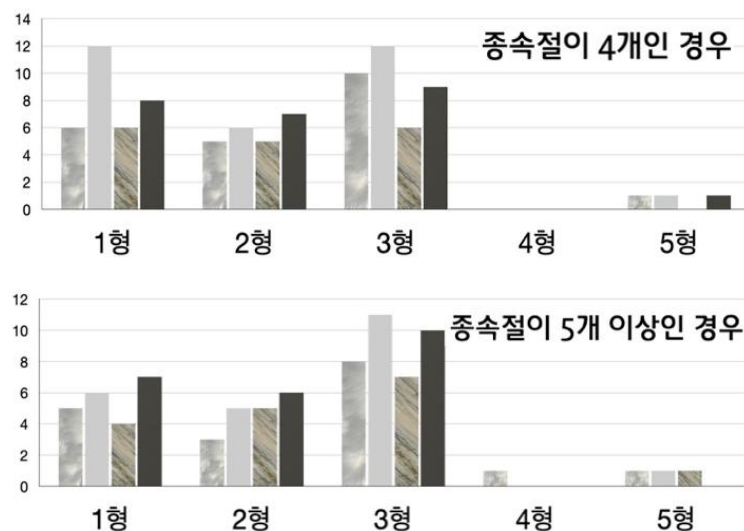
집단	절2개	1형	2형	3형	4형	5형
males young	50	35	22	39	1	2
males old	60	43	24	49	1	3
females young	40	25	21	31	1	2
females old	59	39	26	48	1	3

집단	절3개	1형	2형	3형	4형	5형
males young	16	16	10	19	1	1
males old	19	19	12	23	1	1
females young	14	11	13	17	0	1
females old	21	19	14	27	1	2

집단	절4개	1형	2형	3형	4형	5형
males young	6	6	5	10	0	1
males old	8	12	6	12	0	1
females young	4	6	5	6	0	0
females old	6	8	7	9	0	1

집단	5 이상	1형	2형	3형	4형	5형
males young	3	5	3	8	1	1
males old	4	6	5	11	0	1
females young	3	4	5	7	0	1
females old	4	7	6	10	0	0





〈그림 7〉 종속절의 개수에 따른 집단별 5형식 문형 분포

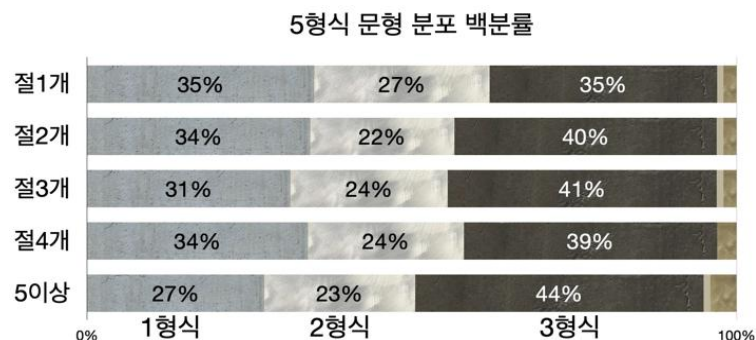
〈표 4〉와 〈그림 7〉에서 보듯이, 종속절의 5형식 문형 분포도 앞서 〈그림 4〉의 주절의 5형식 문형 분포에서와 유사하게 1, 2, 3형식 문형이 대다수를 차지하고 있고, 4, 5형식 문형은 그 빈도가 매우 적음을 발견할 수 있다.

종속절이 1개인 경우에 대하여 종속절의 5형식 문형 분포에 미치는 성별과 연령의 영향을 알아보기 위하여 Mann-Whitney U 검정을 시행한 결과, 젊은 집단에서 성별 요인의 영향은 $W=1,038,672$, $p>0.05$ 로 무의미한 것으로 나타났으며, 마찬가지로 나이 든 집단에서도 $W=1,535,486$, $p>0.05$ 로 성별의 영향이 없었다. 남성 집단 내에서 연령 요인이 미치는 영향은 $W=1,316,831$, $p>0.05$ 로 없었으며, 여성 집단 내에서도 $W=1,190,532$, $p>0.05$ 로 연령 차이에 따른 영향은 없었다. 결과적으로 종속절이 1개인 경우 연령이나 성별의 영향이 없이 주로 1, 2, 3형식 문형이 대다수를 차지하고 있

음을 알 수 있다.

종속절이 2개인 경우에 대하여 검정을 시행해 보면, 젊은 집단과 나이 든 집단 모두에서 각각 $W=394,280$, $W=698,188$, $p>0.05$ 로 성별이 5형식 문형 분포에 영향을 미치지 않았고, 남성과 여성 집단에서도 각각 $W=589,780$, $W=456,172$, $p>0.05$ 로 연령의 영향이 없었다. 종속절이 3개인 경우에는 나이 든 집단에서만 $W=188,646$, $p<0.05$ 로 성별이 5형식 문형 분포에 영향을 미쳤고 나머지 모든 경우에서 영향이 없었고, 종속절이 4개인 경우에는 젊은 집단과 남성 집단에서만 각각 $W=15,473$, $W=28,160$, $p<0.05$ 로 성별과 연령이 5형식 문형 분포에 영향을 미쳤고 나머지는 영향이 없었다. 마지막으로 종속절이 5개 이상인 경우, 성별과 연령이 모든 경우에 영향을 미치지 않는 것으로 나타났다.

종속절의 개수별로 집단을 모두 합쳐 빅아이 코퍼스 전체에 대하여 종속절의 개수에 따른 1, 2, 3형식 문형 각각의 평균 백분률을 계산해 보면 <그림 8>과 같다.



<그림 8> 종속절의 1, 2, 3형식 문형 백분률 (평균)

앞서 <그림 5>에서 주절의 집단별 문형 분포에서는 1형식 < 2형식 < 3형식의 순서로 백분률이 증가하였으나, <그림 8>에서는

이와 다소 다르게 2형식 < 1형식 < 3형식의 순서로 백분율이 증가하는 것을 볼 수 있다. 순서는 다소 바뀌었지만, 전체적으로 1, 2, 3형식이 대다수를 차지하는 경향은 주절과 종속절 모두 마찬가지였다.

4. 결론

본 논문은 우리나라 영문법 교육에서 큰 비중을 차지하고 있는 5형식 문형 구조가 영어 구어체 코퍼스인 벅아이 코퍼스에서 어느 정도의 빈도로 분포하고 있는지를 조사하기 위하여, 벅아이 코퍼스를 구성하고 있는 모든 발화 문장의 구조를 5형식 문형에 근거하여 남녀노소 집단별로 분석하였다. 5형식 문형의 형태를 온전하게 지닌 발화 문장을 가능한 많이 찾아내기 위하여 구어체의 특성을 반영한 여덟 가지의 지침을 만들었고, 이에 따라 음성 코퍼스에서 추출한 개별 발화 문장을 우선 단문, 중문, 복문으로 구분한 뒤, 대다수를 차지하고 있는 단문과 복문의 주절 및 종속절을 5형식 문형으로 분류하여 기술통계와 비모수 추론통계인 Mann-Whitney U 검정으로 남녀노소 집단별로 분석하였다.

그 결과 벅아이 코퍼스의 화자들은 하나의 발화 문장 당 평균 11 개 정도의 단어를 담고 있었으며, 화자 당 평균 6천여 단어로 6백여 문장을 발화한 것으로 나타났다. 발화 문장의 주절의 5형식 문형 분포를 살펴보면, 통계 분석 결과 성별과 연령의 요인이 분포에 대체적으로 영향을 미쳤고, 압도적인 대다수를 차지하는 문형은 1형식, 2형식, 3형식 문형이었으며 4형식과 5형식은 고작 1~3% 정도의 비중을 차지하고 있었다. 1~3형식 문형과 4~5형식 문형의 두 집단이 엄청난 빈도의 차이를 보이는 반면, 1~3형식 집단 내에서 각 문형의 빈도는 1형식과 2형식 문형들이 20% 중후반

대의 대등한 빈도를 보이는 가운데 3형식 문형이 40%대 초중반대의 빈도를 보이는 것으로 나타나, 3형식 문형이 가장 빈도가 높고 애용되는 것으로 밝혀졌다.

복문의 구조를 지닌 발화 문장의 종속절을 그 개수에 따라 구분하여 종속절의 5형식 문형 빈도를 살펴본 결과, 단문과 복문의 비중이 6:4 정도의 비율을 보였으며, 종속절의 경우 그 개수에 상관없이 일부 경우를 제외하고는 대체로 성별과 연령이 큰 영향을 미치지 않는 반면, 주절의 문형 분포에서처럼 1형식, 2형식, 3형식이 압도적인 대다수였고 그 중 3형식이 40%대로 최대 다수를 차지하고 있음이 밝혀졌다.

이번 연구에서 4형식과 5형식 문형의 빈도가 예상 외로 상당히 낮게 나온 것은 흥미롭고 또 의외의 발견으로 볼 수 있는데, 현재와 앞으로의 영문법 문형 구조 교육에 있어서 우리나라 학습자들에게 어떤 의미를 지닐 것인지 고민이 필요할 수 있다고 생각된다. 교육 현장에서 영어 문형 교육의 필수적인 요소로 여겨지는 5형식 문형 교육에 있어서 과연 1~5형식 문형을 동일한 비중으로 교육할 것인지 본 연구와 같은 구어체 문형 분포 조사를 바탕으로 검토가 필요할 것으로 생각된다. 기존의 교육 방식이 교사나 관련 학자들의 주관적 판단에 근거하여 우리말과 다르거나 특이한 문형으로 생각되어질 수 있는 4형식이나 5형식 문형 교육에 교육적 자원을 상대적으로 더 쓰거나, 1~5형식 문형을 동일한 비중으로 가르치는 것은 바람직하지 않을 것으로 생각된다. 영어 교육의 목표가 원활한 의사소통이라면 원어민들이 압도적으로 많이 쓰는 문형에 더욱 많은 시간과 노력을 들이는 것이 합리적일 것이기 때문이다.

참고문헌

- 김진란, 「고등학교 영어 교과서에 나타난 영어 구동사의 코퍼스 기반 분석」, 전남대학교 석사학위논문, 2019.
- 김현주, 「용어색인을 활용한 영어 문법 학습의 효과와 영어 학습에 대한 태도 변화 분석」, 한양대학교 석사학위논문, 2016.
- 배소영, 「학술적 에세이 코퍼스를 기반으로 한 한국 대학생과 영어 원어민 대학생의 관계대명사 which와 that의 사용 비교」, 이화여자대학교 석사학위논문, 2018.
- 배영남, 「5형식 문형 이론과 영어 문법 교육」, 『언어과학연구』 24, 2003, pp. 83-110.
- 서아리, 「영어 주어-동사 도치 구문에 관한 코퍼스 기반 연구」, 계명대학교 박사학위논문, 2022.
- 안효빈, 「한국 영어 학습자 구어 코퍼스 기반 어휘 묶음 사용 양상 분석 연구」, 인천대학교 석사학위논문, 2022.
- 이문복·신동광·전유아·원장호·이희종·강동길, 「코퍼스 활용을 통한 영어교사의 실제적 영어 쓰기지도 능력 향상 방안 연구」, 연구보고서, 한국교육과정평가원, 2008.
- 임수영·이은주, 「코퍼스를 활용한 문법 과제가 고등학생의 영어 문법 학습과 정의적 반응에 미치는 영향」, 『영어학』 12(2), 2012, pp. 303-325.
- 정덕교, 「영어 문장형식의 연구 및 적용 - 영문법 5형식의 재조명」, 『교양교육연구』 4(2), 2010, pp. 177-204.
- 정송휘, 「코퍼스를 활용한 문법 학습이 대학생의 문법 능력에 미치는 영향」, 한남대학교 박사학위논문, 2014.
- 최인지, 「영국 영어 대화 코퍼스에 나타난 담화표지어 like 연구」, 『새한영어영문학』 64(4), 2022, pp. 149-174.
- 한인석, 「코퍼스를 이용한 영어 동의 형용사 연구: “able, capable, competent & qualified”를 중심으로」, 『언어연구』 38(3), 2022, pp. 307-322.
- Boersma, Paul, “Praat, a system for doing phonetics by computer”, *Glot International* 5:9/10, 2001, pp. 341-345.
- Onions, C. T., *An Advanced English Syntax*, London: Keagun Paul, 1932, pp. 4-38.
- Onions, C. T., *Modern English Syntax* (New ed. Of An Advanced English Syntax, by B.D.H. Miller), London: Routledge & Keagun Paul, 1971.

- Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E, *Buckeye Corpus of Conversational Speech* (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor), 2007.
- R Core Team., *R: A language and environment for statistical computing*, *R Foundation for Statistical Computing*, Vienna, Austria. URL <https://www.R-project.org/>, 2020.
- RStudio Team, *RStudio: Integrated development for R*[Computer software], Boston, MA: RStudio, 2019, Retrieved from <http://www.rstudio.com/> on March 20, 2024.

(투고일: 2024. 5. 18 심사완료일: 2024. 6. 20 게재확정일: 2024. 6. 21)

이유리, 윤규철
영남대학교 영어영문학과
[38541] 경북 경산시 대학로 280
kyoon@ynu.ac.kr

[Abstract]

A Study on the English Sentence Structure of the Buckeye Corpus

Lee, Yu-Ri & Yoon, Kyu-Chul
(Yeungnam University)

The purpose of this work was to investigate in the Buckeye corpus of spontaneous English speech the frequency distribution of the five types of English sentence structure popular in the English grammar classes of Korea. All the utterances from the four groups divided by sex and age in the corpus were classified into simple, compound and complex sentences. The complex sentences were further classified depending on the number of subordinate clauses in a sentence. Then the main and subordinate clauses of all the utterances were identified as belonging to one of the five types of sentence structure. The result showed that an overwhelming number of sentences were of the types 1, 2 and 3, with the types 4 and 5 showing very low frequencies. Among the first three types, types 1 and 2 accounted for a similar proportion, whereas type 3 showed the highest frequency, approximately 1.5 times higher than the others. The findings seem to be meaningful in considering how educational resources should be directed in the future English grammar classes in Korea.

Key word: Buckeye corpus, sentences, utterances, sentence structure, sentence analysis, English