코퍼스 기반 어휘 분석 유감

고광윤
(연세대학교)

Goh, Gwang-Youn. 2011. Corpus-based Vocabulary Studies in Korea: Limitations and Suggestions. *The Linguistic Association of Korea Journal. 19*(1). 41-60. This article investigates corpus-based English vocabulary studies carried out in Korea for the last ten years in an attempt to discuss their current status and limitations and suggest ideas for possible solutions and future directions in this research area. To attain this research goal, the present study examines a total of 40 research articles and theses that mainly aim to present an analysis of overall vocabulary level of texts. The results indicate that many of these studies need to compile a more accurate word list and to use more appropriate methods of comparison for more reliable and fruitful results. In particular, virtually no studies seem to have succeeded in providing a systematic analysis of overall vocabulary level of texts, by not going beyond the comparison of the vocabulary of the given texts with certain vocabulary lists or with the vocabulary of other comparable texts. With these limitations in mind, the present study further discusses the properties of two major problems, explores some possible solutions, and suggests the implications of the present study for future directions.

Key Words: corpus, vocabulary study, word list, type, token, TTR, lemma, lemmatization, vocabulary level, lexical coverage.

1. 서론

언어의 교육과 학습에서 한 동안 문법이나 발음에 비해 부차적인 것으로 인식되었던 어휘는 이제 성공적인 언어습득과 효율적인 언어사용을 위해 필요한 가장 핵심적인 요인 가운데 하나로 여겨지고 있다(cf. Wilkins 1972, McCarthy 1990, Willis 1990, Nattinger & DeCarrico 1992, Lewis 1993, Nation 1993, Coady 1997, Carter 1998, DeCarrico 2001). 언어의 습득 및 사용에서 이처럼 매우 큰 비중을 차지하는 어휘에 대한 연구는 지난 20여 년 동안 컴퓨터 기술을 바탕으로 눈부시게 발전한 코퍼스언어학 덕분에 많은 변화와 성

과를 이루어왔다(cf. Meyer 2004, McEnery et al. 2006, O'Keeffe et al. 2007). 이러한 변화의 주목할 만한 내용 가운데 하나는 바로 교육이나 학습을 위한 어휘의 선정이 교육전문가나 원어민의 직관에 의존하기보다는 실제 사용되는 언어자료의 모음인 코퍼스를 바탕으로 과학적이며 객관적으로 이루어져야 한다는 것인데, 이러한 믿음은 이제 더 이상 큰 논란의 여지가 없이 매우 보편적으로 수용되고 있는 듯하다.

언어습득과 사용에 있어서 어휘가 차지하는 이와 같은 핵심적 역할을 반영하여 국내에서도 그 동안 수많은 (영어) 어휘 관련 연구가 진행되어 왔으며, 근래에 이루어진 어휘 관련 연구들은 실증적 연구의 국제적인 흐름에 발맞추어 대부분이 코퍼스 분석을 기반으로 이루어지고 있다. 좀 더 구체적으로, 지난 10년 동안 국내에서는 영어어휘와 이의 습득 및 사용을 직접적으로 취급한 연구들이 총 130건 이상 발견되는데, 이들 영어어휘 관련 연구들은 다양한 영어텍스트의 어휘적 내용과 영어어휘의 학습 및 교수에 대한 우리의 이해에 상당한 기여를 해온 것으로 보인다. 하지만 이들 연구들은, 본 연구의 다음 섹션에서 자세히 다루고 있는 바와 같이, 코퍼스 기반 어휘 연구의 지속적인 발전을 위해 해결해야 할 중요한 문제점들을 동시에 지니고 있으며, 특히 연구결과의 타당성과 신뢰성을 충분히 확보하기 위한 노력이 시급한 것으로 판단된다.

한편, 국내외의 활발한 코퍼스 기반 연구의 흐름과 관련하여 코퍼스 기반 언어 연구나 언어교육 연구의 현황을 분석한 연구는 찾아보기가 매우 어려운 상황이며, 영어교육 분야에서 코퍼스 기반 연구의 동향을 파악하려고 노력한 이은주(2008) 정도가 있을 뿐이다. 더욱이 기존 코퍼스 기반 연구들의 내용을 심도 있게 검토하여 그 한계와 문제점을 파악하고 해결방안이나 구체적인 연구방향을 제시한 연구는 사실상 전무한 실정이다. 따라서 본 연구에서는 그 동안 국내에서 이루어진 코퍼스 기반 영어어휘 연구들의 전체적인 현황을 파악하고 그 내용을 면밀히 검토함으로써 보다 발전적인 연구를 위한 아이디어와 시사점을 제공하고자 한다. 또한 기존 연구들에서 공통적으로 발견되는 몇 가지 중요한 문제점들에 대해서는 가능한 대안을 모색하고 구체적인 사례에 적용함으로써 문제 해결의 단초를 제공하고자 한다. 본 연구에서는 분석의 심도를 높이기 위해 분석의 대상이 되는 연구의 범위를 제한하여 코퍼스 기반 어휘 연구 가운데에서도 코퍼스의 분석과 가장 밀접하게 관련된, 영어텍스트의 어휘수준 및 내용을 분석한 연구들을 집중적으로 다루게 될 것이다.

2. 코퍼스 기반 영어어휘 연구의 현황

2.1. 전반적 연구 동향

국내에서 진행된 코퍼스 기반 어휘 분석 연구의 전반적인 동향을 파악하기 위한 조사대

상의 검색과 수집은 다음과 같이 이루어졌다. 우선, 국내의 주요 학술 데이터베이스인 DBPIA, RISS, KISS와 국회도서관의 상세검색에서 [어휘+영어(+코퍼스)]를 검색어로 사용하여 분석대상이 될 논문을 찾아 정리하였는데, 검색대상은 2001년부터 2010년까지 발표된 모든 국내학술지 게재 논문과 학위논문으로 한정하였다. 이와 같은 1차 조사결과를 정리하면 아래의 표와 같다.[1)]

표 1. (코퍼스 기반) 영어어휘 연구 검색결과

검색 DB	국회도서관	RISS	KISS	DBPIA
검색어	어휘+영어+코퍼스		어휘+영어(+코퍼스)	
기간	2001～2010			
검색결과	134건	135건	61건	97건
학술지	22	22	61(3)	97(3)
학위논문	112	113	-	-

위와 같은 방식으로 얻은 코퍼스 기반 영어어휘 관련 연구들은 크게 다음과 같은 3가지 유형으로 나누어 설명할 수 있다. 첫째, 해당 텍스트의 어휘수준과 내용을 전체적으로 분석한 연구로서, 이들 연구는 주로 텍스트의 전반적인 어휘수준을 분석하여 텍스트들 상호간의 어휘수준을 비교하거나 특정 교재가 지닌 어휘적인 면에서의 연계성 등을 평가하려는 목적을 지니고 있다(예: 권인숙 2002a, 설영 2007, 이영빈 2009; 유성자 2007, 김인숙 2009, 한지영 2010). 둘째, 전체적인 어휘수준이나 내용보다는 주어진 텍스트에 사용된 일부 어휘들의 사용양상과 특징을 질적 또는 양적으로 분석하려 한 연구를 들 수 있는데, 서법조동사를 비롯한 문법적 어휘들의 분포, 다의어의 쓰임, 연어관계(collocation)의 분석 등이 그 주된 내용을 이루고 있다(예: 정은아 외 2008, 문영인 2009, 정연창 외 2009). 마지막으로, 코퍼스의 장점을 어휘학습에 응용한 연구를 들 수 있는데, 이들 연구는 주로 코퍼스를 활용하여 어휘학습의 효과를 알아보거나 학습교재 개발을 목표로 하고 있다(예: 홍선이·오선영 2008, 박소연·윤현숙 2009, 정다혜 2009; 임명숙 2010, 민덕기 외 2010).

이들 코퍼스 기반 영어어휘 관련 연구들의 전반적인 동향에서 발견할 수 있는 한 가지 흥미로운 사실은, 기존 연구의 대부분이 국내의 (교육)대학원에서 학문 후속세대인 새내기 영어

1) 데이터베이스의 규모가 크고 학술지 논문과 학위논문에 대한 정보를 모두 제공하는 국회도서관(http://www.nanet.go.kr)과 RISS(http://www.riss.kr)에서는 [어휘+영어+코퍼스]를 검색어로 사용하였으며, 규모가 상대적으로 작고 학술지 논문만을 포함하는 DBPIA(http://www.dbpia.co.kr/)와 KISS(http://kiss.kstudy.com/)에서는 3가지 검색어를 모두 합하여 검색할 경우 (표의 괄호 안에 주어진 것처럼) 각각 3편의 논문만이 검색되어 우선 [어휘+영어]를 검색한 후 검색결과를 검토하여 실제 분석대상이 될 논문을 선별하였다.

(교육)학도들에 의해 작성된 석사학위논문이란 점이다. 이러한 사실은 무엇보다 한국의 영어학과 영어교육 연구에 실증적 노력이 강하게 반영되고 있다는 면에서 바람직한 모습이라고 할 수 있다. 하지만 이들 학위논문들의 내용을 자세히 들여다보면 안타깝게도 그 주제와 내용이 뚜렷한 이유 없이 중복되고 심지어는 여러 해에 걸쳐 반복되는 경우도 종종 발견되는데(예: 유성자 2007, 우현이 2007, 천윤희 2008, 한지영 2010, 최서희 2010), 이는 개인적 연구노력의 낭비를 초래함은 물론 장기적으로는 한국 영어(교육)학 분야의 발전을 저해하는 요인이 될 수 있으므로 앞으로의 연구에서 특히 유의해야 할 부분이라 여겨진다.

2.2. 어휘수준 분석 연구의 현황과 문제점

2.2.1. 분석대상 및 방법

국내의 코퍼스 기반 어휘 관련 연구들 가운데 본 논문에서 그 내용을 좀 더 심도 있게 고찰하고자 하는 것들은 위에 주어진 분류에서 주로 첫 번째 유형에 속하는 것들로, 주어진 텍스트의 전체적인 어휘수준을 분석한 논문들이다.[2] 이들 심층 분석대상 논문들은 4개의 국내 주요 학술 데이터베이스에서 검색된 모든 관련 논문들의 내용을 면밀히 검토하여 선별한 것으로서 총 40편에 이르며, 학위논문과 학술지 게재 논문이 각각 32편과 8편이다.[3]

이들 논문들의 내용과 어휘수준 분석 연구의 전반적인 현황을 파악하기 위해 먼저 각 연구의 분석대상과 어휘의 분석을 위해 사용한 분석도구,[4] 그리고 정확한 어휘수준 분석의 필수적인 기초라 할 수 있는 레마처리 여부를 파악하였다. 다음은 그 내용을 정리한 표이다.

---

2) 다른 두 유형으로 분류될 수 있는 연구들도 전체적인 어휘수준의 분석을 포함한 것들은 심층 분석대상에 포함시켰다.

3) 본 연구의 심층 분석대상인 연구는 다음과 같다. 학술지논문: 권인숙(2002b, 2004), 정미애(2007), 고광윤·박정준(2007), 김종국(2008), 서은경(2008), 이용훈(2008), 윤현숙(2009); 학위논문: 권인숙(2002a), 안미정(2005), 이진희(2005), 김수정(2006), 문소정(2006), 원수정(2006), 김영미(2007), 설영(2007), 송해영(2007), 우현이(2007), 유성자(2007), 최사라(2007), 허미영(2007), 강상요(2008), 김영지(2008), 김은정(2008), 김현진(2008), 이지영(2008), 이하림(2008), 주헌우(2008), 천윤희(2008), 최수경(2008), 김소라(2009), 김인숙(2009), 이영빈(2009), 이혜원(2009), 강문선(2010), 이연경(2010), 이자연(2010), 임명숙(2010), 최은주(2010), 한지영(2010).

4) 이들 분석도구 가운데 비교적 널리 쓰이는 다른 도구와 달리 FreCol은 단어빈도 조사와 연어추출을 위해 해당 연구자가 Visual Basic을 사용해 제작했다는 프로그램임(권인숙 2002a,b). 그 외 모든 분석도구에 대한 설명은 다음을 참조할 것: AntConc (http://www.antlab.sci.waseda.ac.jp/antconc_index.html), Concordance (http://www. concordancesoftware.co.uk/), NLPTools (이용훈 2007), Range (http://www. victoria.ac.nz/lals/staff/paul-nation.aspx), SCP = Simple Concordance Program (http://www.textworld.com/scp/), WS = WordSmith Tools (http://lexically.net/).

표 2. 논문의 유형 및 연구 방법

구분	유형별 논문 수
논문종류	학위논문 32편(석사: 31, 박사: 1), 학술지논문 8편
분석대상	교과서(26), 대입시험(5), 신문(2), 읽기책(2), 독해교재(1), 성경(1), 비즈니스영어(1), 대학교양영어(1), 대학생 발화(1)
분석도구	WS(23), SCP(5), NLPTools(4), Concordance(3), FreCol(3), AntConc(2), Range(2)[5]
레마처리	자동처리(11), 수작업(5),[6] 무처리(23), 미상(1)[7]

위의 표에 요약 정리된 분류 내용에서 엿볼 수 있는 것처럼 코퍼스 기반 어휘수준 분석 연구들은 다음과 같은 몇 가지 특징을 지닌다. 먼저, 논문의 유형에서는 학위논문, 그 중에서도 일종의 훈련과 연습의 성격을 지니고 있다고 할 수 있는 석사학위논문이 약 80%를 차지하고 있는 것으로 나타났다. 앞으로도 후학들의 지속적인 관심과 연구노력을 격려해야 하겠으나 동시에 어휘수준 분석에 대한 중견학자들의 좀 더 많은 관심과 본격적인 연구가 기대된다고 하겠다.

둘째, 분석의 대상으로 삼은 텍스트 자료는 영어교과서가 과반수를 웃도는 65%를 차지하고 있어 분석대상의 중복으로 인한 연구노력의 낭비가 우려된다. 앞으로는 (특히 학위논문을 위한 연구에서) 새로운 연구목적과 뚜렷한 이유를 가진 경우가 아니라면 가급적 영어교과서의 분석을 지양하고 교과서 이외의 학습자료를 선택하거나 영화대본, 문학작품, 영문 홈페이지 등 기타 다양한 영어텍스트의 분석을 권장하고 적극적으로 시도할 필요가 있겠다.

셋째, 코퍼스 분석도구(corpus tools)로는 WordSmith Tools가 가장 널리 쓰이고 있었다. 이는 해당 프로그램이 제공하는 기능의 다양성과 일반적인 지명도 및 신뢰성을 고려할 때 자연스런 결과라고 생각된다. 하지만 유료인 WordSmith Tools 못지않게 충분히 강력하고 다양한 기능을 제공하면서도 무료로 사용할 수 있는 프로그램과 더불어 어휘수준 분석을 위해 특화되어 사용이 편리한 프로그램도 있으므로(예: AntConc, Range) 코퍼스 기반 연구의 저변확대와 보다 효율적인 연구를 위해 WordSmith Tools 이외의 다른 다양한 코퍼스

5) 분석도구에서 전체 합계(42)가 분석논문 수(40)보다 많은 이유는 이연경(2010)과 주헌우(2008)가 각각 두 개의 프로그램(WS+SCP, WS+Range)을 사용했기 때문이다.

6) 수작업으로 레마처리를 했다고 밝히고 있는 연구는 권인숙(2002a, 2002b, 2004), 원수정(2006), 주헌우(2008)이다.

7) 레마처리 여부를 명시적으로 밝히지 않은 일부 논문의 경우에는 논문의 내용을 바탕으로 연구자가 레마처리 여부를 평가하였다. 한편, 정미애(2007)의 경우 레마처리 여부에 대한 명시적인 언급이 없을 뿐 아니라 연구의 내용을 살펴보아도 레마처리 여부를 판단하기 어려워 미상으로 처리하였다.

분석도구의 사용을 고려해보는 것도 유익하리라 생각된다.

마지막으로, 전체 연구들의 약 58% 가량이 정확한 어휘규모의 산정을 위해 필요한 레마처리작업(lemmatization)을 행하지 않고 어휘수준의 분석을 시도한 것으로 밝혀졌다. 또한 레마처리를 실시한 연구들도 거의 예외 없이 정확성과 일관성이 결여될 수밖에 없는 수작업에 의존했거나 (WordSmith Tools와 같이 일부 코퍼스 분석 프로그램에서 제공하는) 불완전한 자동레마처리 기능을 사용하는 데 그치고 있었다. 레마처리는 각 단어의 변이형들을 기본형에 통합시켜(예: *forget* ← *forgets, forgot, forgotten, forgetting*) 어휘의 규모를 정확하게 파악하도록 도와주는 작업으로서 객관적인 어휘수준 분석을 위해 반드시 필요한 필수적인 기초 작업이라 할 수 있다. 본 연구의 3절에서 자세히 논의하고 있는 바와 같이, 레마처리를 생략하면 어휘의 규모가 터무니없을 정도로 부풀려지게 되며, 자동적인 레마처리를 실시할 경우에도 (완벽한 처리가 사실상 불가능하기 때문에) 분석의 성격과 방식에 따라 중요한 결과의 차이를 보일 수 있다. 따라서 레마처리와 관련된 문제들에 대한 연구자들의 각별한 주의와 관심이 요구된다고 하겠다.

2.2.2. 분석 내용

한편, 어휘의 전체적인 수준을 분석한 연구들의 분석내용을 좀 더 자세히 살펴보기 위해 각 연구에서 취급한 구체적인 분석항목을 살펴본 결과는 다음과 같다.[8)]

표 3. 분석 내용

분석 내용	논문 수
어휘의 규모(토큰, 타입, TTR)	37
고빈도 어휘의 빈도	17
특정 어휘목록과의 비교	16
연어(collocation)	9
신출어휘	9
유형별(주요 품사나 내용어) 분포	6
다의어의 사용양상	3
어휘다발(lexical chunks)	1
키워드	1

8) 분석내용을 정리한 <표 3>에서 2번째 열의 논문 수는 총 40편의 분석대상 논문 가운데 각각의 분석내용을 다룬 논문의 수를 가리킨다.

위 표에 정리된 분석결과에서 어느 정도 짐작할 수 있는 것처럼, 지난 10년간 이루어진 어휘의 수준과 내용 분석을 위한 국내 연구는 크게 두 가지 유형으로 대별할 수 있다. 우선, 가장 전형적인 유형으로서 분석대상인 텍스트에 대해 토큰(token), 타입(type), TTR(type-toke ratio),[9] 고빈도 어휘의 빈도 등, 사용된 전체 어휘의 규모에 대한 기초적인 통계자료를 산출한 후, 이를 바탕으로 서로 다른 텍스트를 비교하거나 교육과학기술부 지정 기본어휘목록이나 GSL(General Service List)과 같은 특정 어휘목록에 포함된 어휘들의 반영률을 알아 본 연구들을 들 수 있다(예: 안미정 2005, 김영지 2008, 이용훈 2008, 이영빈 2009).[10] 또 하나의 일반적인 유형은 분석대상 텍스트에 쓰인 어휘의 전체적인 규모를 알아본 후 (많은 경우 이와 무관하게) 소수의 고빈도 어휘나 특정 그룹에 속하는 일부 어휘들의 분포와 사용 양상을 살펴 본 연구들이다(예: 권인숙 2002b, 2004, 정미애 2007, 서은경 2008, 윤현숙 2009).

분석 내용의 다양성과 차이에도 불구하고 주어진 텍스트에 쓰인 어휘의 수준을 알아보는 것이 핵심적인 부분이라 할 수 있는 이들 연구들은 나름대로의 장점과 적지 않은 기여에도 불구하고 해결해야 할 몇 가지 중요한 문제점을 공통적으로 지니고 있다. 그 가운데 가장 두드러진 것은 아마도 지금까지의 관련 연구들이 대부분 어휘의 수준을 분석함에 있어서 어휘의 단순한 외형적 크기를 비교하는 데 그치고 있다는 점일 것이다. 즉, 기존의 어휘 분석 연구들은 사실상 거의 예외 없이 토큰, 타입, TTR 등을 산출하여 제시하거나 이를 바탕으로 서로 관련된 자료를 비교하고 특정 어휘목록의 반영률을 알아보는 정도에 그치고 있어 주어진 텍스트에 쓰인 어휘의 전체적인 수준이나 내용 구성을 충분히 밝히지 못하고 있는 것이다. 이는 기존의 연구들이 지니고 있는 방법론적인 문제점, 특히 정확한 어휘목록의 작성문제와 더불어 앞으로의 보다 나은 연구를 위해 개선이 시급한 부분이라고 사료된다.

정리하자면, 지금까지 국내에서 이루어진 어휘의 수준 분석을 위한 연구들은 그 심각성의 정도는 다르지만 거의 예외 없이 (보다 방법론적인 면에서) 정확한 어휘목록 작성과 (보다 내

9) 토큰은 사용된 모든 단어의 전체 출현 수(즉, 한 단어가 10번 나오면 10개로 계산), 타입은 중복된 경우를 모두 하나로 계산한 서로 다른 단어의 수를 가리킨다. 한편, TTR은 어휘밀도(lexical density), 즉 일정한 단위의 텍스트 내에 얼마나 많은 다른 어휘가 쓰이고 있는지를 보여주는 지표인데, 전체 타입 수를 토큰 수로 나눈 후 100을 곱하여 산출한다.

10) GSL은 현재 가장 널리 알려진 빈도 기반 기초어휘목록 가운데 하나로서 본래 West(1953)가 Lorge & Thorndike(1938)을 참고하여 선정한 2,000개의 어휘군(word family)으로 출발했으며, 이후 여러 차례에 걸쳐 수정 보완을 거쳤는데, 지금은 2,284개의 단어로 구성된 Bauman & Culligan(1995)의 개정판이 흔히 쓰인다. 이밖에 기존 어휘 분석 연구에서 어휘의 반영률 산출에 자주 활용되어 온 어휘목록으로는 Collins Cobuild 사전의 기초어휘목록, Oxford 대학출판부에서 2005년 Oxford Advanced Learner's Dictionary(7판)와 함께 만든 기본어휘목록인 Oxford 3000(cf. Hornby & Wehmeier 2005), 기초필수 학술어휘 570개로 구성된 AWL(Academic Word List, Coxhead 2000, Nation 2001) 등이 있다.

용적인 면에서) 실질적인 어휘수준 분석이라는 두 가지 문제점을 공통적으로 지니고 있다고 할 수 있다. 따라서 본 연구의 나머지 부분에서는 구체적인 어휘 분석 사례를 통해 이러한 두 가지 문제의 속성과 관련 내용을 좀 더 상세히 알아본 후 문제 해결을 위한 아이디어를 제시하고 적용해봄으로써 보다 발전된 어휘 분석 연구를 위한 지속적인 논의를 촉발하고자 한다.

3. 어휘 분석의 문제점과 대안모색

3.1. 코퍼스와 분석도구

본 절에서는 영어텍스트에 쓰인 어휘의 전체적인 규모와 수준을 분석하는 과정에서 노출된 두 가지 문제점과 그 의미를 구체적인 사례 분석을 통해 좀 더 상세히 살펴보고 해결방안을 모색하여 대안을 제시해 보고자 한다. 이를 위해 선택한 자료는 한국의 대입영어시험과 관련 모의고사들인데, 좀 더 구체적으로 그 대상은 제7차 교육과정에 속하는 2004년부터 2009년까지 실시된 대학수학능력시험(수능, 11월)과 평가원의 대학수능모의평가(모의, 6, 9월) 및 교육청의 전국연합학력평가(학력, 3, 4, 7, 10월)의 영어시험 텍스트이다. 유사한 수준과 내용을 지니고 있다고 알려진 이들 대입 관련 영어시험은 연중 실시 횟수가 서로 다르기 때문에 텍스트의 규모를 통일하기 위해 지난 6년간 시행된 세 종류의 시험에서 해마다 각각 1회분씩의 영어시험 텍스트를 모아 코퍼스로 구축하였다.

표 4. 대입 영어시험 코퍼스

시험	주관기관	구축대상	기간
대학수학능력시험	한국교육과정평가원	11월 영어시험	2004년 ~2009년
대수능모의평가	한국교육과정평가원	9월 영어시험	
전국연합학력평가	교육청	10월 영어시험	

한편, 본 연구에서 어휘목록을 작성하고 레마처리를 시행함으로써 어휘의 전체적인 규모를 파악하고 수준을 분석하기 위해 사용한 코퍼스 분석도구는 현재 국내외를 막론하고 가장 널리 쓰이고 있는 WordSmith Tools 5.0이다. WordSmith Tools가 제공하는 핵심적인 기능은 크게 [Concord], [WordList], [KeyWord]의 세 가지로 나눌 수 있는데, 본 연구의 사례분석 과정에서는 어휘목록을 작성해 주고 레마처리 작업과 구성 어휘의 비교 및 편집 작업 등을 효율적으로 할 수 있도록 해주는 [WordList]와 그 하위기능들을 주로 사용하였다.

3.2. 어휘목록의 작성

먼저, 어휘의 규모와 수준을 파악하기 위해 필요한 사전 작업으로서 모든 어휘 분석의 기초라고 할 수 있는 어휘목록의 작성 문제를 살펴보자. 하나 혹은 일련의 텍스트에 쓰인 모든 어휘들의 목록 작성은 대부분의 코퍼스 분석도구에서 원하는 텍스트 파일을 탑재한 후 해당 명령을 실행하면 아주 손쉽게 이루어진다. 아래의 표는 WordSmith Tools의 [WordList] 기능을 사용하여 대입 관련 영어시험 코퍼스에 쓰인 모든 어휘들의 목록을 만든 후 이를 바탕으로 전체 어휘의 규모를 나타내는 기초적인 지표를 정리한 것이다.

표 5. 대입 관련 영어시험의 어휘규모 (레마처리 전)

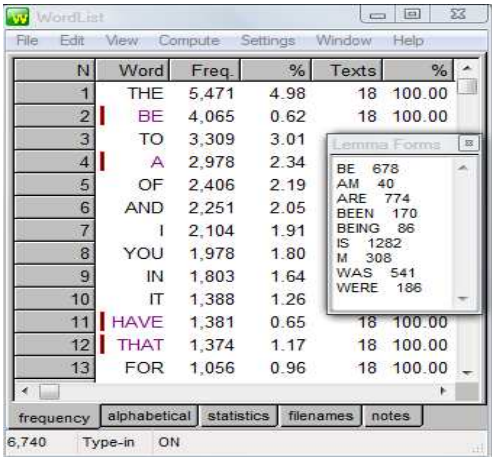
시험	수능	모의	학력	통합
타입	5,409	5,215	5,381	9,376
토큰	36,620	35,730	36,904	109,254
TTR	14.77	14.60	14.58	8.58

위의 표에 주어진 지표 가운데 특히 타입 수를 보면 각 영어시험의 어휘규모를 어느 정도 짐작할 수 있는데, 약간의 차이가 있으나 모두 비슷한 수준이며 세 시험을 모두 합할 경우 약 9,400단어 수준에 이름을 알 수 있다. 이러한 분석결과는 그 자체만을 토대로 판단할 때 한국의 대입영어시험이 약 9,000 단어 수준이라는 의미로 해석될 수도 있을 것이다. 하지만 그러한 해석은 여러 가지 면에서 타당성이 결여된 것이라 할 수 있는데, 그 가장 큰 이유는 무엇보다 각 단어의 변이형들을 모두 해당 기본형에 통합시켜주는 레마처리를 거치고 나면 어휘의 수가 매우 큰 규모로 줄어들기 때문이다.

WordSmith Tools에서는 미리 준비된 레마목록(lemma list) 파일을 탑재하여 레마처리를 자동으로 실행할 수 있게 해주는 기능이 있다.[11] 이러한 과정을 거치면 아래의 그림에서 보는 바와 같이 단어의 변이형들을 모두 해당 기본형에 통합해주어 훨씬 실제에 근접한 어휘규모를 파악할 수 있게 될 뿐 아니라, 각 기본형에 속한 모든 변이형들의 빈도에 대한 정보까지 함께 제공하므로 매우 유용한 기능이다.

11) 레마목록은 각 단어의 변이형과 기본형을 하나의 그룹으로 묶어 정리한 목록인데 현재 가장 널리 쓰이는 것은 Y. Someya가 1998년에 작성한 것이다(http://www.lexically.net/downloads/e_lemma.zip). 이 레마목록은 40,569개의 단어형(토큰)을 14,762개의 그룹으로 분류하여 정리한 것으로 매우 포괄적이긴 하지만 동일한 형태를 지닌 여러 단어가 존재하는 문제 등으로 인해 완벽한 레마처리는 어려우며, 사용자가 목적에 따라 편집 가공하여 활용하는 것이 바람직하다.

그림 1. 레마처리 결과



그렇다면, 레마처리를 거칠 경우 각 대입 관련 영어 시험의 어휘규모가 얼마나 달라질까? 다음은 위에 설명한 것과 같은 방식으로 자동화된 레마처리(file-based joining)를 실행한 후 그 결과를 정리한 것이다.

표 6. 대입 관련 영어시험의 어휘규모 (레마처리 전후)

시험		수능	모의	학력	통합
타입	전	5,409	5,215	5,381	9,376
	후	4,069	3,942	4,093	6,740
	차이 (비율)	1,340 (32.9)	1,273 (32.3)	1,288 (31.5)	2,636 (39.1)
토큰		36,620	35,730	36,904	109,254
TTR		11.11	11.03	11.09	6.20

위의 표에 주어져 있는 것처럼, 레마처리 과정을 거친 결과 각 시험에서 어휘의 수가 대략 1,300개 정도씩 감소했으며, 세 시험을 모두 통합하여 분석한 경우에는 약 2,600 단어 정도 감소한 것으로 나타났다. 이러한 차이는 레마처리 후의 (실제에 가까운) 어휘규모를 기준으로 계산할 때, 개별시험의 경우 실제 어휘 수보다 약 30% 이상씩, 통합시험의 경우엔 약 40%가량이나 부풀려 진 규모이다. 이러한 결과는 무엇보다 우리가 실제 연구수행 과정에서 레마처리 없이 어휘의 규모나 수준을 분석하고 이를 바탕으로 서로 다른 텍스트를 비교할 경

우 얼마나 터무니없는 결과를 낳게 될지 매우 분명하게 보여주는 것이라 할 것이다.[12]

그렇다면, 현재 이용 가능한 코퍼스 분석도구에서 제공하는 레마처리 기능을 활용하면 사용된 모든 단어의 기본형들만으로 구성된 완벽한 어휘목록을 얻을 수 있을까?[13] 안타깝게도 이 질문에 대한 답변은 다소 부정적이다. 즉, 주어진 레마처리 기능을 활용하여 레마처리 작업을 행한다 할지라도 현재 제공되는 프로그램의 기능만으로는 모든 단어에 대한 레마처리가 자동으로 완벽하게 이루어지지는 않기 때문에 정확한 어휘규모의 산정이 결코 쉽지 않은 것이다. 이와 같이 완벽한 레마처리가 어려운 이유는 무엇보다 주어진 텍스트 안에서 어떤 단어가 그 기본형은 발견되지 않고 변이형(들)만 쓰이고 있을 경우 (레마처리 과정에서 해당 변이형들을 통합시킬 기본형이 존재하지 않아) 컴퓨터 프로그램이 그 변이형들을 모두 별도의 독립된 레마로 간주하여 어휘목록을 작성하기 때문이다. 예를 들어, 어떤 텍스트에 *concentrate*란 단어의 기본형은 존재하지 않고 굴절형인 *concentrates, concentrated, concentrating*만 쓰이고 있을 경우 3개의 굴절형이 모두 독립된 레마로 처리되고 결과적으로 서로 다른 타입으로 간주되게 되어 정확한 어휘의 규모 산출에 문제가 발생하는 것이다.

본 연구에서는 기본형 부재 등의 이유로 레마처리가 제대로 이루어지지 않은 단어들의 규모를 파악하기 위해 실제로 자동 레마처리를 통해 산출된 어휘목록에서 기본형이 아닌 단어들을 모두 찾아 일일이 수정하였다. 수정을 거친 어휘목록을 바탕으로 산출된 각 대입 관련 영어시험과 통합 텍스트의 (타입 기준) 어휘규모는 아래와 같다.

표 7. 대입 관련 영어시험의 어휘규모 (타입) (수정 전후)

시험	수능	모의	학력	통합
전	4,069	3,942	4,093	6,740
후	3,947	3,799	3,948	6,459
차이 (비율)	122 (3.1)	143 (3.8)	145 (3.7)	281 (4.4)

12) 영어시험과 유형이 다른 소설의 경우에도 그 결과가 크게 다르지 않은데, 해리포터(Harry Potter) 7권 모두의 텍스트를 코퍼스로 구축하여 타입을 산출한 결과 레마처리 전후가 각각 23,988개와 17,648개로서 레마처리를 하지 않을 경우 실제 규모보다 약 36% 가량 더 과장된 결과를 낳는 것으로 나타났다.

13) 어휘의 규모를 산출하기 위해 단어를 분류하는 방식은 크게 사전적인 표제어를 기본형(주로 동사의 원형 부정사형이나 명사의 단수형)으로 하여 그 굴절형들을 한 그룹으로 묶는, 보다 일반적인 레마(lemma) 방식(예: *allow* ← *allows, allowed, allowing*)과 굴절어는 물론 관련 파생어까지도 다수 포함시켜 하나의 대표어(head word) 아래 묶어 처리하는 어휘군(word family) 방식(예: *allow* ← *allows, allowed, allowing, allowance, allowable*)이 있다. 어떤 기준과 방식으로 단어들을 분류하여 어휘목록을 작성해야 하는가 하는 문제는 고려해야 할 다양한 요인들로 인해 적지 않은 논란이 수반되는 매우 흥미로운 연구주제이다(cf. Bauer & Nation 1993, Heatly et. al. 2002).

위의 표에 정리된 수정 전후의 타입 수를 살펴보면 자동 레마처리를 거친 후에도 불완전한 처리로 인해 실제 어휘규모보다 약 3~4.5% 정도 부풀려져 있음을 알 수 있다. 혹자는 불완전하게 처리된 부분의 비율이 비교적 작게 느껴져 어쩌면 자동적인 레마처리만으로도 충분하다고 생각할지 모른다. 이러한 바람은 어쩌면 당연한 것이라 느껴지는데, 더 완벽한 어휘목록을 얻기 위해서는 자동으로 레마처리가 완료된 1차적인 어휘목록을 보면서 불완전하게 처리된 부분을 모두 찾아내 일일이 수작업을 통해 수정하는 상당히 지루한 과정을 거쳐야 하기 때문이다.[14] 하지만 문제는 일견 작아 보일지도 모르는 이러한 차이가 연구자가 다루고자 하는 문제가 무엇이냐에 따라 실제로는 매우 중요한 역할을 할 수도 있다는 데 있다. 특히, 다루는 문제가 특정 어휘의 출현빈도 자체와 밀접한 관련을 지닐 때 일견 작아 보일 수 있는 차이가 통계학적으로 매우 중요한 차이를 가져올 수도 있으므로 그 차이의 통계학적인 중요성이나 의미는 연구주제와 취급문제의 속성에 따라 신중히 판단하여 이해되어야 할 것이다.

불완전한 레마처리로 인한 차이의 통계학적인 의미 해석 문제와 더불어 좀 더 완벽한 레마처리가 필요한 또 하나의 이유는 연구자가 분석하려는 텍스트의 규모에 따라 수정 전후의 차이가 더 커질 수 있기 때문이다. 특히, 코퍼스의 규모가 작을수록, 처리가 제대로 되지 않는 단어들의 비중이 커지는 경향이 있는데, 실제로 2009년도 수능영어시험만을 가지고 통계를 산출한 결과 자동 레마처리만을 거친 경우와 이후 수정작업을 거친 후의 어휘규모가 타입 수를 기준으로 각각 1,538 단어와 1,464 단어로 나타나, 자동 레마처리만을 거칠 경우 실제 규모보다 약 5.1% 정도 더 과장된 결과를 보였다. 이와 관련하여 한 가지 주목해야 할 것은, 본 논문의 2절에서 이미 살펴본 바와 같이, 기존 관련 연구 가운데 (소수의) 영어교과서를 분석대상으로 삼은 경우가 상당히 많을 뿐 아니라 코퍼스의 규모가 비교적 크더라도 한 코퍼스 안의 여러 소규모 텍스트들을 상호 비교하는 경우가 매우 흔하다는 점이다. 따라서 설사 (자동) 레마처리를 거친다 할지라도 산출된 어휘의 규모가 실제보다 상당히 과장될 가능성이 있으며, 또한 그런 통계를 바탕으로 한 비교나 분석 결과는 충분한 타당성을 지닌다고 보기는 어려울 것이다.

요약하자면, 어휘의 전체적인 규모와 수준을 분석함에 있어서 충분한 타당성을 지닌 연구결과를 원한다면 레마처리는 필수적인 작업이라 할 수 있다. 또한 많은 경우 자동 레마처리 작업을 거친 결과도 결론의 타당성에 부정적인 영향을 미칠 수 있으므로 가급적이면 추가적인 수정작업을 통해 좀 더 정확한 어휘목록을 얻도록 노력해야 할 것이다.

---

14) 컴퓨터 프로그래밍 능력이 없는 일반 연구자들의 경우 현재의 여건에서 이러한 수정보완 작업을 좀 더 용이하게 할 수 있는 작은 지혜 가운데 하나는 자동 레마처리된 어휘목록을 역순정렬([Edit]-[Other sorts]-[Reverse word] 선택)하여 단어들을 굴절어미의 형태별로 한 곳에 모아 편집하는 것이다.

3.3. 어휘의 수준 분석

주어진 영어텍스트에 쓰인 어휘의 규모나 수준의 분석을 목표로 하거나 그 연구내용이 이와 밀접하게 관련된 연구들이 공통적으로 지니고 있는 또 하나의 큰 아쉬움은 지금까지의 어떤 국내 연구에서도 전체적인 어휘수준에 대한 충분히 포괄적이고 타당성 있는 분석을 찾아보기 어렵다는 것이다. 즉, 본 논문의 2.2.2에서 논의한 바와 같이 지금까지의 관련 연구들은 대부분 텍스트의 어휘수준을 알아보기 위해 토큰과 타입 등의 기초적인 통계수치를 산출하여 제시하고, 이를 바탕으로 서로 다른 텍스트를 비교하거나 (널리 알려진 어휘목록에 기반을 둔) 어휘들의 반영률을 알아보는 것에서 크게 벗어나지 못하고 있다.

물론 기초어휘목록 등 특정 목적의 어휘목록에 포함된 어휘들의 반영률을 통해 주어진 텍스트의 어휘적 성격을 어느 정도는 파악할 수 있을 것이다.[15] 하지만, 어휘목록의 반영률이나 일부 어휘의 사용양상을 분석함으로써 해당 텍스트에 사용된 전체 어휘들의 수준과 구성 내용을 제대로 파악할 수 있는지 의문이다. 특히, 어떤 영어텍스트에 대해 우리가 정말 알고 싶은 것 가운데 하나는 그 텍스트에 쓰인 어휘의 전체적인 수준, 즉, 주어진 텍스트의 충분한 이해를 위해 필요한 어휘의 양과 그 수준이라고 할 수 있는데, 지금까지 일반적으로 행해진 방식의 분석을 통해서는 이와 같은 실제적인 어휘수준을 파악하기가 쉽지 않은 것이다.

좀 더 구체적으로, 우리가 어휘목록 작성의 문제를 논의하기 위해 살펴 본 대입 관련 영어시험의 경우를 생각해보자. 이 경우 우리가 전체적인 어휘수준과 관련하여 답해야 하는 문제는 "이 영어시험의 실제적인 어휘수준은 어느 정도인가?", 즉 "이 영어시험에 충분히 대비하기 위해 대입수험생들이 학습해야 할 영어어휘의 양과 난이도는 어느 정도인가?" 라고 할 수 있을 것이다. 이러한 질문에 대해 혹자는 대입 관련 영어시험에 사용된 모든 어휘의 수를 계산하면 이 시험의 어휘수준을 알 수 있지 않겠느냐고 말할지 모른다. 하지만 이런 방식으로는 대입영어시험의 어휘수준을 제대로 파악하기가 어려운데, 그 이유는 시험의 횟수가 누적될수록 사용된 어휘의 수가 (적어도 어느 정도까지는 계속) 증가하게 될 것이기 때문이다.

그렇다면, 일반 연구자가 손쉽게 특정 텍스트에 쓰인 전체 어휘의 수준을 파악하려면 어떻게 해야 할까? 이러한 문제를 해결하기 위해 필자가 본 연구에서 제안하고 싶은 방법은 대규모 코퍼스를 바탕으로 일정한 단위로 누적된 일련의 어휘목록을 만든 후 그 어휘수준을 분석하고자 하는 텍스트에 사용된 어휘와 비교하여 전체 어휘의 수준별 구성을 알아보는 것이다. 이제 그러한 방법에 대해 좀 더 자세하게 살펴보도록 하자.

---

15) 하지만 그런 경우에도 단지 주로 빈도를 바탕으로 선정된 기본적인 어휘의 반영률만을 살펴볼 것이 아니라 더 나아가 분석대상 텍스트의 성격에 알맞은 적절한 유형의 어휘목록을 신중하게 선택하여 분명한 목적을 가지고 비교하는 것이 해당 텍스트의 어휘수준이나 성격을 알아보기 위해 더 바람직할 것이다. 구체적인 예를 들자면, 수능영어시험의 경우 사용된 어휘를 AWL 등의 기초학술어휘목록과 비교하고 반영률을 살펴봄으로써 대학수학능력 평가시험으로서의 적합성, 즉, 이 영어시험이 대학에서의 공부에 필요한 학생들의 영어능력을 평가하는 데 얼마나 적합한가를 평가해볼 수 있을 것이다.

먼저, 텍스트에 사용된 전체 어휘의 수준별 구성을 분석하기 위해 가장 대표적인 대규모 영어코퍼스 가운데 하나인 BNC(British National Corpus)의 어휘목록을 레마처리한 후 빈도순으로 2,000 단어씩 계속 누적시켜 40,000 단어 수준에 이르는 총 20개의 어휘목록을 만들어 준비하였다. 그 다음, BNC의 수준별 어휘목록 각각과 대입 관련 영어시험의 통합된 어휘목록을 비교함으로써 BNC의 각 어휘수준에서 영어시험에 사용된 어휘들이 전체의 몇 퍼센트나 포함되어 있는지 계산하였다. 이러한 어휘비교를 위해 본 연구에서는 WordSmith Tools에서 제공하는 단어의 매칭기능을 활용하였는데, 이 기능을 사용하면 두 목록에 포함된 어휘들의 일치 여부와 그 정도를 확인할 수 있어 매우 유용하다. 다음은 이와 같은 방식으로 대입영어시험 통합 텍스트에 쓰인 전체 어휘의 수준별 구성을 분석하여 정리한 표이다.

표 8. 대입영어시험 어휘의 수준별 구성 (BNC 기반)[16]

어휘목록 / 수준	레마처리 전	레마처리 후	수정 후
BNC 2000	1,778 (18.96)	1,778 (26.38)	1,822 (28.21)
BNC 4000	3,001 (32.01)	3,001 (44.53)	3,213 (49.74)
:	:	:	:
BNC 20000	5,095 (54.34)	5,095 (75.60)	5,668 (87.75)
BNC 22000	5,165 (55.09)	5,165 (76.63)	5,747 (88.98)
BNC 24000	5,214 (55.61)	5,214 (77.36)	5,806 (89.89)
BNC 26000	5,260 (56.10)	5,260 (78.04)	5,856 (90.66)
:	:	:	:
BNC 38000	5,430 (57.91)	5,430 (80.56)	6,040 (93.51)
BNC 40000	5,444 (58.06)	5,444 (80.77)	6,057 (93.78)

위의 표를 살펴보면, 영어시험의 통합 텍스트를 바탕으로 작성된 세 개의 각 대입영어 어

16) 표의 각 셀에 주어진 숫자는 BNC의 각 수준에 포함된 단어의 수와 그 비율을 가리킨다. 한편 레마처리 전과 후의 어휘목록이 보여주는 수준별 어휘 수가 동일한 이유는 자동 레마처리 과정에서 변이형들이 해당 기본형에 통합되어 어휘의 전체 타입 수는 감소하지만 기본형의 수와 내용은 변화가 없기 때문이다.

휘목록들에 있는 단어들이 BNC 기반의 각 어휘수준에서 어느 정도의 비율로 포함되어 있는지를 알 수 있다. 또한, 이를 통해 우리는 어휘의 외형적인 규모만으로는 짐작하기 어려운 (해당 텍스트에 사용된) 어휘의 전체적인 수준을 어렵지 않게 파악할 수 있다.

여기에서 한 가지 매우 흥미로운 사실은 어휘목록의 작성 방법에 따라 차이를 보이는 어휘수준의 분석결과인데, 레마처리를 하고 수정까지 거친 목록의 경우 24,000 단어 수준에서 약 90% 정도의 어휘가 설명되는 반면,[17] 레마처리를 거치지 않은 어휘목록과 레마처리 후 수정을 거치지 않은 목록의 경우에는 동일한 수준에서 각각 56%와 77% 정도의 어휘만이 설명된다는 점이다. 이러한 분석결과의 차이를 통해 우리는 무엇보다 레마처리를 하지 않은 상태에서 어휘수준을 분석하려 할 경우 얼마나 터무니없는 결과가 나올 수 있는지 다시 한 번 분명하게 확인할 수 있다. 또한, 경우에 따라서는 레마처리된 목록의 수정을 통해 생긴 작은 차이가 결과적으로 매우 중요한 차이를 가져 올 수 있다는 점에도 유의해야 것이다.

4. 결론

지금까지 본 연구에서는 지난 10년간 국내에서 이루어진 코퍼스 기반 어휘 분석 연구들의 현황을 고찰하고 이를 통해 발견한 중요한 문제점들과 그 해결방안에 대해 논의하였다. 지금까지의 논의를 정리하면 다음과 같다.

먼저, 수많은 코퍼스 기반 어휘 분석 연구들의 대부분을 차지하는 것은 석사학위논문인데, 학문 후속세대들의 실증적 연구에 대한 관심은 매우 고무적인 반면, 연구 주제와 내용이 중복되는 경우가 많고 특히 영어교과서의 분석이 지나치게 높은 비중을 차지하여 연구노력의 낭비가 우려된다. 앞으로는 교과서 이외의 다양한 영어자료에 대한 분석과 새로운 문제의 취급에 관심을 기울일 필요가 있다.

둘째, 어휘수준 분석을 시도한 연구들 가운데 약 60% 정도가 어휘목록을 작성할 때 레마처리 작업을 생략함으로써 정확한 어휘규모의 산출이 매우 어렵게 되는 문제점이 발견되었다. 레마처리를 생략하면 대부분의 경우 어휘의 규모가 지나치게 부풀려져 터무니없는 분석결과를 낳게 되므로 각별한 주의가 요망된다. 또한, 레마처리를 거친 목록도 좀 더 정확하고 객관적인 분석결과를 얻기 원한다면 불완전하게 처리된 부분에 대해 별도의 수정보완 작업을 실시하는 것이 바람직하다. 하지만 이러한 작업은 매우 지루하고 시간도 많이 소요되므로 관

17) 이러한 사실은 (만일 단순하게 전체 어휘의 90% 정도를 적절한 어휘학습량이라 가정할 경우) 대입영어 관련시험이 실제로 24,000단어 수준이라는 뜻은 아니다. 무엇보다 레마처리와 수정을 거친 대입영어시험 어휘목록에는 인명이나 지명 등의 고유명사와 고교수준을 넘어서는 1회성 단어들이 상당 수 포함되어 있기 때문에 실제 어휘수준은 더 낮을 것이라 생각된다. 참고로 대입영어시험 어휘의 약 80%는 약 12,000 단어 수준인 것으로 분석되었다.

련 작업의 대부분을 자동화할 수 있는 알고리듬의 개발이 절실히 요구된다고 하겠다.

셋째, 대부분의 연구들이 타입, 토큰 등의 기초통계를 산출하고 이를 바탕으로 서로 다른 텍스트를 비교하거나 고빈도 어휘 등 일부 어휘의 사용양상을 살펴보고 특정 어휘목록의 반영률을 살펴보는 데 그침으로써 사용된 어휘의 전체적인 수준을 분석하는 데에는 미흡한 것으로 드러났다. 이러한 문제점을 해결하기 위한 시도로서 본 연구에서는 BNC 등의 대규모 영어코퍼스를 바탕으로 일련의 수준별 어휘목록을 작성하고 분석대상 텍스트에 쓰인 어휘들의 수준별 구성 비율을 산출하여 평가하는 방법을 제시하였다. 앞으로 어휘수준의 분석을 위한 다양한 방법들이 활발하게 논의되기를 기대한다.

마지막으로, 코퍼스 기반의 어휘 연구에 대한 현황 파악을 통해 드러난 문제점들 가운데 완벽하게 레마처리된 정확한 어휘목록의 작성 문제는 그 해결이 결코 쉽지 않으나 모든 다른 분석을 위한 기초가 되므로 연구의 발전을 위해 시급히 해결되어야 할 중요한 문제이다. 코퍼스 기반의 실증적 언어 연구에 관심을 가진 다양한 분야의 연구자들이 서로 협력하여 자동 레마처리 후의 수정작업이 훨씬 더 효율적으로 진행될 수 있도록 관련 프로그램 상의 보완이 가까운 시일 내에 이루어 질 수 있게 되기를 기대한다.

참고문헌

강문선. 2010. 코퍼스를 기반으로 한 대학수학능력시험 외국어 영역 듣기 평가 어휘 분석: 2005학년도부터 2009학년도 중심으로. 석사학위논문. 상명대학교 교육대학원.

강상요. 2008. 고등학교 영어교과서에 나타난 동사+명사구 연어 형태의 코퍼스 기반 연구. 석사학위논문. 상명대학교 교육대학원.

고광윤・박정준. 2007. 중학교 영어 교과서의 어휘적 연계성에 대한 코퍼스 바탕 연구. *영어학연구, 24,* 27-45.

권인숙. 2002a. 韓國, 日本, 中國 중학교 영어 교과서의 코퍼스 언어학적 어휘 비교 분석. 박사학위논문. 숭실대학교 대학원.

권인숙. 2002b. 중학교 영어 교과서의 코퍼스 언어학적 어휘 비교 분석. *영어교육, 57*(4), 409-444.

권인숙. 2004. 한국 중학교 6차 및 7차 교육과정 영어교과서의 코퍼스 언어학적 어휘 비교 분석. *Foreign Languages Education, 11*(1), 211-251.

김소라. 2009. 코퍼스 기반 아동용 영자신문 어휘 분석과 연어 표현집 개발. 석사학위논문. 청주교육대학교 교육대학원.

김수정. 2006. 중학교 영어교과서 어휘의 코퍼스 분석: 양상 조동사를 중심으로. 석사학위논문. 연세대학교 교육대학원.

김영미. 2007. 말뭉치를 통한 고등학교 영어교과서의 어휘 분석. 석사학위논문. 연세대학교 대학원.

김영지. 2008. 대학수학능력시험 영어 어휘수준에 대한 코퍼스 바탕 연구. 석사학위논문. 연세대학교 교육대학원.

김은정. 2008. 고등학교 교과서 양상조동사의 코퍼스 언어학적 분석. 석사학위논문. 연세대학교 교육대학원.

김인숙. 2009. 코퍼스를 이용한 영자신문 어휘 분석연구. 석사학위논문. 중앙대학교 교육대학원.

김종국. 2008. A corpus-based investigation into Korean learners' spoken vocabulary use. *영어영문학연구, 50*(4), 65-84.

김현진. 2008. 영어교과서의 연계성에 관한 코퍼스 바탕 연구: 중학교 3학년과 고등학교 1학년을 중심으로. 석사학위논문. 연세대학교 교육대학원.

문소정. 2006. 코퍼스 언어학적 접근법에 의한 중등학교 영어 교과서의 어휘 분석 연구. 석사학위논문. 한국교원대학교 대학원.

문영인. 2009. '요청'에 나타나는 우리나라 중학생들의 어휘 사용에 대한 코퍼스적 분석. *영어교육연구,* 21(3), 203-223.

민덕기 · 김소라 · 김규화. 2010. 아동용 영자 신문 활용 코퍼스 기반 영어 연어집 개발에 관한 연구. 멀*티미디어 언어교육, 13*(1), 173-200.

박소연 · 윤현숙. 2009. 코퍼스를 활용한 영어어휘 학습의 효과 연구. *영어영문학연구, 51*(3), 145-165.

서은경. 2008. 한국과 싱가포르 초등학교 영어 교과서 어휘 비교 분석. *영어교육연구, 20*(3), 225-249.

설영. 2007. EFL 환경과 ESL 환경의 중학교 영어교과서 어휘비교: 한국과 홍콩의 대표적인 교과서 중심으로. 석사학위논문. 서울시립대학교 대학원.

송해영. 2007. 고등학교 영어 교과서에 나타난 어휘의 코퍼스 언어학적 분석. 석사학위논문. 전남대학교 교육대학원.

안미정. 2005. 한국과 싱가포르 영어교과서의 코퍼스 언어학적 어휘 비교분석. 석사학위논문. 성균관대학교 교육대학원.

우현이. 2007. 코퍼스에 근거한 초등학교 6학년과 중학교 1학년 영어교과서 어휘의 연계성 분석. 석사학위논문. 한국외국어대학교 교육대학원.

원수정. 2006. 코퍼스에 기반한 중학교 교과서 어휘 분석. 석사학위논문. 인하대학교 교육대학원.

유성자. 2007. 영어교과서의 연계성에 관한 코퍼스 바탕 연구: 초등학교 6학년과 중학교 1학년을 중심으로. 석사학위논문. 연세대학교 교육대학원.

윤현숙. 2009. 개정 교육과정 중학교 1학년 영어교과서에 나타난 어휘의 코퍼스 기반 분석. *현대영어교육, 10*(2), 87-107.

이연경. 2010. 초등영어교과교육을 위한 단계별 읽기 책 어휘의 코퍼스 언어학적 분석. 석사학위논문. 중앙대학교 교육대학원.

이영빈. 2009. 한국, 중국, 일본 대입 영어시험 어휘수준에 대한 코퍼스 바탕 연구. 석사학위논문. 연세대학교 교육대학원.

이용훈. 2007. *NLPTools를 이용한 코퍼스 분석과 활용: 언어학 연구, 영어교육, 그리고 영어 교재 개발에서의 활용*. 서울: 케임브리지.

이용훈. 2008. 코퍼스 분석기법을 이용한 대학교 교양언어 교재 어휘목록 분석의 실제. *인문학연구, 35*(2), 125-149.

이은주. 2008. 영어교육과 응용언어학 분야에서 수행된 코퍼스 기반 연구의 분석. *영어교육 63*(2), 208-306.

이자연. 2010. 2007년 개정 교육과정에 따른 중학교 1학년 영어교과서 어휘의 코퍼스 분석. 석사학위논문. 한국외국어대학교 교육대학원.

이지영. 2008. 영어성경의 복음서에 활용된 어휘의 코퍼스 기반 분석. 석사학위논문. 전남대학교 대학원.

이진희. 2005. 중학교 영어 교과서의 다 빈도 어휘와 언어형태의 양적 코퍼스 분석. 석사학위논문. 전남대학교 교육대학원.

이하림. 2008. 중학교 3학년을 위한 영어읽기교재의 코퍼스 분석. 석사학위논문. 숙명여자대학교 교육대학원.

이혜원. 2009. 코퍼스에 기반한 제 7차 교육과정과 개정 교육과정의 중학교 1학년 영어교과서 어휘 비교 분석. 석사학위논문. 충남대학교 교육대학원.

임명숙. 2010. 코퍼스 기반 영어 교과서 신출어휘 분석과 어휘 자료집 개발. 석사학위논문. 한국교원대학교 대학원.

정다혜. 2009. 고등학교 영어학습자의 코퍼스 활용 학습의 효율성 연구. *교과교육연구, 2*(2), 431-461.

정미애. 2007. 코퍼스 분석을 통해 살펴본 비즈니스 영어의 어휘적 특성. *언어, 32*(4), 751-777.

정연창 · 고은정 · 김은일. 2009. 한국 대학생의 영어 '경동사+명사' 연어 능력에 관한 코퍼스 기반 연구. *언어과학, 16*(2), 61-81.

정은아 · 윤은순 · 이용훈. 2008. 중 · 고등학교 영어교과서에 나타난 조동사 Can의 분포에 대한 코퍼스언어학적 연구. *Foreign Languages Education, 15*(1), 363-382.

주헌우. 2008. A Corpus-based analysis of vocabulary in the BEWL and the CSAT. 석사학위논문. 고려대학교 대학원.

천윤희. 2008. 코퍼스 언어학적 분석을 통한 초중등 영어 교과서의 연계성 연구: 초등학교 6학년과 중학교 1학년 교과서를 대상으로. 석사학위논문. 한국교원대학교 교육대학원.

최사라. 2007. Are authorized English textbooks in authentic English for Korean middle school students? A corpus study on conversation sections. 석사학위논문. 연세대학교 대학원.

최서희. 2010. 콘코댄서를 활용한 초·중등 영어 교과서의 연계성 분석. 석사학위논문. 경인교육대학교 교육대학원.

최수경. 2008. 영어 리더스북의 어휘수준에 대한 코퍼스 기반 연구. 석사학위논문. 연세대학교 교육대학원.

최은주. 2010. 고등학교 영어교과서와 대학수학능력시험 외국어(영어)영역 어휘수준에 대한 코퍼스 바탕 분석. 석사학위논문. 연세대학교 교육대학원.

한지영. 2010. 코퍼스에 근거한 초등학교와 중학교 영어 교과서의 연계성 분석. 석사학위논문. 한양대학교 교육대학원.

허미영. 2007. 영어 독해교재의 코퍼스 바탕 연구: 어휘 분석을 중심으로. 석사학위논문. 연세대학교 교육대학원.

홍선이 · 오선영. 2008. 한국 고등학생들을 대상으로 한 코퍼스 기반 어휘 및 문법 학습의 효과. *영어교육연구, 20*(1), 261-283.

Bauer, L. and Nation, I.S.P. 1993. Word families. *International Journal of Lexicography, 6*(4), 253-279.

Bauman, J. and B. Culligan. 1995. About the General Service List. Retrieved from http://jbauman.com/aboutgsl.html.

Carter, R. 1998. *Vocabulary: Applied Linguistic Perspective*. London: Routledge.

Coady, J. 1997. L2 vocabulary acquisition: A synthesis of the research. In J. Coady & T. Huckin (eds.) *Second Vocabulary Acquisition: A Rationale for Pedagogy* (pp. 273-288). Cambridge: Cambridge University Press.

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

DeCarrico, J. 2001. Vocabulary learning and teaching. In Celce-Murcia, M. (ed.) *Teaching English as a Second or Foreign Language*. 3rd Edition (pp. 285-300). Boston: Heinle & Heinle.

Heatley, A., I.S.P. Nation, and A. Coxhead. 2002. RANGE and FREQUENCY programs. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

Hornby, A. S. and S. Wehmeier. 2005. *Oxford Advanced Learner's Dictionary*. 7th Edition. Oxford: Oxford University Press.

Lewis, M. 1993. *The Lexical Approach: The State of ELT and the Way Forward*. Hove, England: Language Teaching Publications.

Lorge, I. and E. L. Thorndike. 1938. *A Semantic Count of English Words*. Teachers College, Columbia University, New York.

McCarthy, M. 1990. *Vocabulary*. Oxford: Oxford University Press.

McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Meyer, C. F. 2004. *English Corpus Linguistics*. Cambridge: Cambridge University Press.

Nation, I.S.P. 1993. Vocabulary size, growth, and use. In R. Schreuder and B. Weltens. (eds.) *The Bilingual Lexicon* (pp. 115-134). Amsterdam: John Benjamins.

Nation, I.S.P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nattinger, J. and J. DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

O'Keeffe, A., M. McCarthy, and R. Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

West, M. 1953. *A General Service List of English Words*. London: Longman.

Wilkins, D. 1972. *Linguistics and Language Teaching*. London: Edward Arnold.

Willis, J. 1990. *The Lexical Syllabus*. London: Collins.

고광윤
연세대학교 문과대학 영어영문학과
120-749 서울시 서대문구 신촌동 134
전화: 02-2123-2328
이메일: goh@yonsei.ac.kr

Received: 30 November, 2010
Revised: 15 February, 2010
Accepted: 14 March, 2011