키클러스터 분석을 통한 학술목적영어 코퍼스의 어휘다발 특성 연구

장세은 · 이성민*
(한국해양대학교 · 부경대학교)

Jhang, Se-Eun & Lee, Sung-min. (2012). Key clusters analyses of lexical bundles used in English for academic purposes: The Biomed Corpus. *The Linguistic Association of Korea Journal, 20*(4), 219-239. This study is to investigate structural and functional features of English lexical bundles in the Biomed Corpus as a Life Science Corpus through key clusters analyses. Previous studies have identified lexical bundles using their different criteria, but in this study lexical bundles were extracted and identified in a limited number of frequent phrases in comparison to a reference corpus using key clusters analyses. Structural and functional classification of lexical bundles was made following recent work by corpus linguists such as Biber and Barbieri. We found that the lexical bundles that we dealt with represent both positive and negative keyness in value in the list of key clusters, just like keywords analyses. The article examines and discusses the distribution of the frequence category of lexical bundles, the ratio of type and token of nouns in the structures of 'the+Noun+of the' and 'in the+Noun+of', various types of passive constructions, and epistemic stance expressions standing out in functional distribution of lexical bundles in comparison to a reference corpus.

주제어(Key Words): 코퍼스(corpus), 어휘다발(lexical bundles), 키클러스터(key clusters), 긍정적이고 부정적인 핵심도(positive and negative keyness), 키워드(keywords), 생명공학코퍼스(Biomed), 미국영어코퍼스(AmE06)

1. 머리말

20세기말부터 컴퓨터 기술의 발달에 힘입어 코퍼스언어학에 사용되는 프로그램의 툴

* 장세은: 제1저자, 이성민: 제2 및 교신저자

(tools)에서 자동으로 추출할 수 있는 반복적으로 나타나는 단어의 연속체인 빈출 어구(frequent phrases)에 관한 연구가 활발해졌다. 빈출 어구는 Renouf and Sinclair(1991)에서는 연어 틀(collocational frameworks), Biber et al.(1999)에서는 어휘다발(lexical bundles), Scott(1999)에서는 클러스터(clusters), Wray(2000)에서는 정형화된 연속체(formulaic sequences), Granger and Meunier(2008)에서는 어구 연쇄(phraseology) 등으로 불리어 학자마다 같은 의미로 사용되는 말이 다양한 용어로 사용되고 왔는데[1] 최근에는 학술목적영어(English for Academic Purposes, 이하 EAP)에서 사용되는 어휘다발의 특성을 파악하는 연구가 활발히 활발하게 진행되고 있다(Biber et al., 2004; Cortes, 2004; Hyland, 2008; Chen & Baker, 2010; Ädel & Erman, 2012).

이러한 코퍼스연구들은 어휘다발을 대상으로 각 학술장르별로 특징적으로 사용되는 어구 또는 절의 구조적인 분석과 기능적인 분포에 대하여 논의하고 있다. 그러나 이러한 모든 선행연구는 고빈도 분포를 갖는 어휘다발에 대한 연구라서 어떤 특정의 코퍼스, 예를 들어 EAP 또는 특수목적영어(English for Specific Purposes, 이하 ESP)에서만 사용되는 어휘다발에 대한 특성을 살펴보기에는 한계를 갖고 있다. 이를 극복하기 위하여 어휘다발 연구에 키클러스터(key clusters) 분석방법을 사용한 선행연구로는 특정 영국소설가의 문체를 연구한 Mahlberg(2007)와 L1과 L2의 소규모 영문초록 코퍼스로 비교 연구한 Jhang and Hong(2012)[2]이 있다. 특히 어떤 특정의 백만어절 규모의 EAP 코퍼스가 갖는 어휘다발의 특징을 살펴보려면 참조코퍼스와의 대조분석방법인 키클러스터 분석이 매우 유용하다.

본 연구에서는 키클러스터 어휘다발을 추출하기 위하여 워드스미스(WordSmith Tools 6.0) 프로그램을 사용하고, 생명과학저널 코퍼스와 현대 미국문어코퍼스의 어휘다발 목록을 각각 생성한 후 생명과학저널만이 가지고 있는 어휘다발의 구조적, 기능적 특성을 대조 분석으로 살펴볼 것이다. 본 연구의 구성은 다음과 같다. 2장에서는 본 연구의 새로운 연구방법인 키클러스터 분석을 이해하기 위한 필요한 개념과 선행연구에 대해 살펴본다. 3장에서는 연구 방법과 절차를 소개한다. 4장에서는 어휘다발의 구조적 분석과 기능적 분포의 결과를 제시하면서 두드러지게 나타난 주요 특성을 논의한다. 5장에서는 결론부분으로 본 연구의 주요 결과를 요약한다.

2. 키워드와 키클러스터 개념과 선행연구

키워드(keywords)와 키클러스터(key clusters)는 워드스미스 키워드 툴에서 추출하므

1) 본 연구에서는 혼동을 피하기 위하여 빈출 어구를 어휘다발이라는 용어로 통일하여 쓴다.
2) Mahlberg(2007)과 Jhang and Hong(2012)의 주요 연구 내용은 다음 소절의 키클러스터 선행연구에서 언급한다.

로 동일한 추출 메카니즘(mechanism)을 사용하지만 키워드는 단어(word) 단위에서 분석하는 용어이고 키클러스터는 어구(phrase) 단위에서 분석하는 용어이다. 키워드와 키클러스터라는 용어는 하나의 텍스트나 여러 텍스트의 집합으로 구축된 목표코퍼스를 참조코퍼스에 대조하여 산출된 특이한 빈도로 발생하는 단어 또는 어구를 각각 일컫는 말이다.[3] 워드스미스에서는 로그-라이클리후드(log-likelihood) 방법을 사용하여 목표코퍼스의 어휘목록(wordlist)과 참조코퍼스의 어휘목록을 통계 처리하여 키워드를 산출하게 된다. 이렇게 산출된 키워드목록(keywordlist)에는 참조코퍼스보다 많이 사용(over-use)되었을 경우 플러스, 적게 사용(under-use)되었을 경우 마이너스로 표시하여 두 가지 핵심도(keyness)의 값을 보여준다. 키워드목록은 목표텍스트에서 다루고 있는 고유명사나 텍스트의 주제와 관련된 단어를 보여준다. 그러므로 키워드 분석은 분석하고자 하는 텍스트에서 주된 의미영역이 무엇인지 객관적으로 보여줌으로써 직관적으로 파악하지 못했던 사실을 파악할 수 있는 하나의 유용한 분석방법이라 할 수 있다.

위와 같은 키워드의 개념은 Scott(1999)에서 통계수식으로 체계화하였으며 키워드 개념을 활용한 선행연구로는 영어방언 연구(Leech & Fallon, 1992; Oakes & Farrow, 2007), 영어교육 연구(Sardinha & Shimazumi, 2003; Scott & Tribble, 2006), ESP의 해사영어 연구(장세은・변현정, 2011; Jhang & Parent, 2011), 신문담화분석 연구(Scott, 2005; 김혜영・강범모, 2011), 성별, 연령, 지역, 교육 등의 사회언어학 연구(Rayson et al., 1997; 전지은, 2010), 영문학 작품 비교연구(Tribble, 2000; Culpeper, 2002, 2009; Starcke, 2009), 토론 비교연구(Baker, 2004, 2006), 정치학 연구(Johnson et al., 2003; Rayson, 2008) 등이 있다. 이러한 키워드 분석을 통한 연구는 개별 텍스트의 특성을 파악했던 기존의 연구방법과는 달리 다양한 학문영역에서 장르별 또는 다른 텍스트간의 비교를 통해서 분석하고자 하는 텍스트나 코퍼스의 특성을 잘 파악할 수 있는 장점이 있다.

앞서 언급하였지만 키클러스터는 키워드와 동일한 원리로 산출되지만 어휘 단위가 아니라 어구의 단위에서 산출된다는 차이점이 있다. 키클러스터는 목표코퍼스와 참조코퍼스의 단어 간의 빈도수와 코퍼스의 크기간의 통계식에 의해 추출되며 워드스미스의 키워드 툴 기능에서 키워드와 같은 메카니즘을 가진다. 이렇게 생성된 키클러스터목록을 통하여 목표코퍼스 또는 목표텍스트만이 갖고 있는 구문적인 특성을 살펴볼 수 있다. 여기서 주의할 점은 키클러스터와 클러스터의 용어 사용이다. 워드목록 툴, 키워드 툴, 콘코드 툴의 하단 메뉴항목에 있는 클러스터와 본 논문에서 분석하고자 하는 키클러스터와는 통계식에서 많은 차이가 있으므로 용어 사용에 주의를 요한다. 예를 들어, 본 연구의 목표코퍼스의 워드목록, 키워드목록에서 상위 50위내에 공통으로 들어 있는 명사인 *expression, sequence, analysis*를 콘코드 툴의 하단 클러스터 메뉴에서 네 개 단어의 어휘다발을 추출하면 각각 13개, 9개, 8개이지만

3) Scott(1999)은 "A key word, as identified in WordSmith Tools, is a word (or word cluster) which is found to occur with unusual frequency in a given text or set of texts."라고 정의한다.

이 모든 어휘다발의 합인 30개는 본 연구에서 추출한 네 개 단어의 키클러스터 어휘다발 180개에 들어 있지 않기 때문에 우리는 클러스터와 키클러스터에 대한 용어사용에 주의를 해야 한다. 즉, 키클러스터 분석을 통해 추출된 어휘다발과 워드목록과 키워드목록에서 추출된 어휘다발과는 통계식이 상당히 차이가 있다는 것을 알 수 있다.

키클러스터 분석을 통한 어휘다발에 관한 특성을 연구한 선행연구는 Mahlberg(2007)와 Jhang and Hong(2012)이 있다. Mahlberg(2007)은 450만 단어로 구성한 목표코퍼스인 Dickens 코퍼스와 19세기 작가의 작품으로 구성한 450만 단어의 참조코퍼스를 대상으로 다섯 개 단어의 어휘다발 66개의 플러스 핵심도 값과 일곱 개의 마이너스 핵심도 값을 갖는 키클러스터목록을 산출하여 Dickens의 작품이 갖는 문체적 특징을 설명하였다. Jhang and Hong(2012)은 L1과 L2의 영문초록코퍼스의 비교연구로 영어화자의 학자들이 쓴 영문학과 영어학의 SSCI급 국제저널의 초록과 한국인화자의 학자들이 쓴 영문학과 영어학의 국내 한국연구재단등재지 저널의 초록코퍼스를 각각 구축한 후 키클러스터 분석을 통하여 영문학과 영어학 초록에 사용되는 네 개 단어의 어휘다발에 대한 특징을 논의하였다. 도출된 키클러스의 숫자가 영문학초록코퍼스보다 영어학초록코퍼스에 4배 이상 많이 도출되고 영어학초록코퍼스에서 키클러스터 어휘다발은 마이너스 핵심도 값을 갖는 유일한 것이 *the ways in which*로 보고하고 있다.

이러한 키클러스터 어휘다발과 관련된 선행연구를 참고하여 백만어절 규모의 EAP 코퍼스 중에서 영어 어휘와 어구가 정형적인 패턴을 사용할 것으로 예측되는 실험관련 전문학술지인 생명공학 코퍼스인 Biomed 코퍼스를 선정하여 어구 단위인 어휘다발의 특성을 살펴보려면 비슷한 크기의 참조코퍼스와의 대조 분석방법을 통한 키클러스터 분석이 매우 유용할 것으로 본다.

3. 분석 방법

3.1 분석대상과 통계정보

본 연구의 목표코퍼스인 생명공학 코퍼스는 총 1천 4백만 단어로 구성된 미국영어를 대표하는 ANC(American National Corpus)의 하위 부류인 Biomed 코퍼스이다. Biomed 코퍼스는 생물학과 의학저널을 대상으로 2000년에서 2003년까지 4년간 ANC의 연구원들이 구축한 코퍼스이다. 100만 단어 규모의 참조코퍼스인 AmE06 코퍼스와 비교하기 위하여 본 연구에서는 기존의 Biomed 코퍼스에서 279개 텍스트를 새로 선정하여 100만 단어 규모의 목표코퍼스를 새롭게 구축하였으며, 전체 단어 수는 1,092,804개이고 텍스트 수는 279개, 평균 단어 수는 3,916개이다.

참조코퍼스인 AmE06 코퍼스는 Brown Corpus계열에서 최근에 만들어진 현대 미국 문어를 대표하는 일반영어 코퍼스이다. AmE06 코퍼스는 2005년부터 2007년까지 사용된 문어체 미국영어를 바탕으로 구축한 코퍼스로 100만 단어로 이루어져 있으며 모두 15가지 장르의 500개 텍스트로 구성되며 각 텍스트 파일은 약 2,000개 단어가 들어가 있다. AmE06 코퍼스는 장르별로 크게 신문, 소설, 일반산문, 학술산문으로 나누어져 있다.

아래 그림 1은 워드스미스 툴을 사용하여 얻은 목표코퍼스인 Biomed 코퍼스와 참조코퍼스인 AmE06 코퍼스의 통계적 분석이다. 여기서 흥미로운 것은 목표코퍼스와 참조코퍼스의 서로 다른 어휘나 어구 수(타입, type)와 그들의 출현 수(토큰, token)의 비율인 TTR 값의 비교이다.

100만 단어 biomed_files.lst

File Edit View Compute Settings Windows Help

N	text file	file size	tokens (running words) in text	tokens used for word list	sum of ϵ (distinct \ ratio	types	type/token	standardised TTR
1	Overall	11,656,693	1,191,295	1,092,804	26,610	2.44	34.41	
2	1468-6708-3-1.txt	49,088	3,386	3,191	693	21.72	32.07	
3	1468-6708-3-10.txt	60,236	3,780	3,522	883	25.07	34.30	
4	1468-6708-3-3.txt	34,358	2,193	2,059	583	28.31	36.90	
5	1468-6708-3-4.txt	63,818	4,293	4,146	961	23.18	38.05	
6	1468-6708-3-7.txt	36,864	2,325	2,178	650	29.84	36.35	
7	1471-2091-2-10.txt	48,666	3,150	2,634	751	28.51	34.73	
8	1471-2091-2-11.txt	71,384	4,764	4,370	980	22.43	33.17	
9	1471-2091-2-12.txt	50,674	3,307	2,913	816	28.01	36.13	
10	1471-2091-2-13.txt	57,046	3,618	3,412	936	27.43	35.43	

AmE06-TxtFiles_files.lst

File Edit View Compute Settings Windows Help

N	text file	file size	tokens (running words) in text	tokens used for word list	sum of ϵ (distinct \ ratio	types	type/token	standardised TTR
1	Overall	12,187,584	1,014,477	1,000,805	46,600	4.66	45.23	
2	AmE06_A01.txt	24,206	2,032	2,004	826	41.22	47.40	
3	AmE06_A02.txt	24,330	2,030	2,007	857	42.70	48.05	
4	AmE06_A03.txt	25,648	2,029	2,001	788	39.38	45.30	
5	AmE06_A04.txt	26,000	2,053	1,991	742	37.27	42.95	
6	AmE06_A05.txt	24,540	2,014	1,972	764	38.74	43.45	
7	AmE06_A06.txt	25,376	2,030	1,978	804	40.65	48.30	
8	AmE06_A07.txt	25,082	2,010	1,977	768	38.85	43.60	
9	AmE06_A08.txt	24,800	2,036	2,013	777	38.60	43.55	
10	AmE06_A09.txt	25,266	2,026	1,981	775	39.12	44.20	

그림 1. Biomed 코퍼스와 AmE06 코퍼스의 통계정보

TTR의 값을 통해 코퍼스에 사용된 어휘나 어구의 다양성을 측정할 수 있다(Rizzo, 2010). 즉, TTR의 값이 상대적으로 높으면 어휘나 어구가 훨씬 다양하게 사용되고 있다는 것을 의미한다. 위의 그림 1에서 Biomed 코퍼스는 TTR이 2.44, 표준 TTR이 34.41이고 AmE06

코퍼스의 TTR은 4.66, 표준 TTR이 45.23으로 나타났으므로 목표코퍼스인 Biomed 코퍼스가 일반영어인 참조코퍼스 AmE06보다 어휘나 어구의 다양성이 훨씬 더 낮다는 것을 의미한다. 이러한 통계적인 정보는 Biomed 코퍼스는 학문적으로 전문성이 있는 학술분야의 코퍼스이므로 전문적인 용어들이 구체적인 개념이나 설명 또는 과학적인 관찰에 의한 기술 등에서 반복적으로 많이 사용되었다는 것을 의미한다.

3.2 분석 도구 및 과정

본 연구에서 사용한 코퍼스 분석용 프로그램은 워드스미스 툴 6.0이며, 분석한 과정은 다음과 같다. 첫째, 목표코퍼스와 참조코퍼스의 색인(index)[4]목록을 만든다. 이러한 목록을 만드는 과정은 워드스미스의 워드목록 툴에 있는 상단 메뉴의 File에서 New를 클릭한 후 Choose Texts Now에서 각 코퍼스의 전체 텍스트파일을 끌어서 오른쪽 Files selected에 놓고 상단 오른쪽에 있는 OK를 클릭한 후 Make/Add to index를 클릭하면 자동으로 연구자 컴퓨터의 워드스미스 폴더에 자동 저장되는 것이다. 둘째, 각 코퍼스의 어휘다발을 생성한다. 이를 위해서는 만들어진 각각의 색인목록을 워드목록 툴에 있는 상단 메뉴의 File에서 Open을 클릭한 후 지정된 파일에서 각 색인목록을 선택하면 색인목록이 열린다. 상단 Compute메뉴의 Clusters기능을 클릭하여 어휘다발 크기를 네 개 단어로 설정하고[5] 최소빈도를 5회[6]로 정한 후 OK를 클릭하면 네 개 단어 어휘다발이 생성된다. 차후 연구를 위하여 어휘다발을 지정된 폴더에 저장하면 편리하게 사용할 수 있다. 셋째, 키클러스터목록을 생성한다. 이를 생성하기 위해서는 앞서 만들어진 폴더 내의 두 개 코퍼스의 어휘다발목록을 키워드 툴에 있는 상단 메뉴의 File에서 New를 클릭한 후 상단 박스에 있는 목표코퍼스의 어휘다발목록을 탑재하고 아래 박스에는 참조코퍼스의 어휘다발목록을 탑재한 후 Make a key word list now를 클릭하면 키클러스터목록이 자동으로 생성된다.

---

4) 색인은 텍스트에 있는 모든 개별 단어의 위치를 파악하여 개별 문장의 길이나 텍스트의 길이 또는 어휘다발 추출을 위한 클러스터 기능을 사용할 수 있게 해 주는 기능이다(Scott, 2012).

5) 네 개 단어로 이루어진 어휘다발을 연구 대상으로 선정한 이유는 다음과 같다. 첫째, 네 개보다 적은 두 개, 세 개의 단어로 구성된 어휘다발은 문법적인 구조가 명확하게 드러나지 않아 구문적인 분류가 어렵다. 둘째, 네 개보다 많은 다섯 개 여섯 개 단어로 구성된 어휘다발은 길이가 너무 길어지면 코퍼스 유형에 따라 상대적으로 차이는 있겠으나 생명과학의 EAP 또는 해사영어의 ESP와 같은 특수한 코퍼스 유형에서는 전체 추출되는 타입의 숫자가 너무 적어 목표코퍼스의 구문적인 특징을 알아보는 것이 어렵고 고유명사나 내용과 관련된 특정 명사와의 의미적 결합이 강해져서 본 연구의 목적인 담화 기능상의 특징을 살펴보기도 어렵다.

6) 5회 이상으로 설정한 이유는 최소한 5회 이상 빈도가 나타나야 Sinclair(1991)가 주장한 관용의 원리에 적합한 어휘다발로 볼 수 있으며 5회 미만으로 너무 제한하면 연구하고자 하는 코퍼스의 전반적인 특성을 파악하기가 어렵다고 한다.

4. 분석결과와 토의

4.1 어휘다발목록과 키클러스터목록

워드스미스에서 생성된 네 개 단어로 구성된 어휘다발목록에는 Biomed 코퍼스에서 4,273개와 AmE06 코퍼스에서 1,171개의 어휘다발이 있다. 이러한 정보는 아래 그림 2에 주어져 있다. 아래 그림 2는 각 코퍼스의 상위 20위와 마지막 하위 두 개의 어휘다발의 예와 빈도를 보여주고 있다.

원어민100만단어버전_index_4-word clusters.lst

File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%
1	IN THE PRESENCE OF	230	0.02	83	29.75
2	IN THE ABSENCE OF	183	0.02	83	29.75
3	HAS BEEN SHOWN TO	120	0.01	79	28.32
4	IT IS POSSIBLE THAT	93		65	23.30
5	ON THE OTHER HAND	88		52	18.64
6	MM TRIS HCL PH	78		37	13.26
7	AND APPROVED THE FINAL	75		74	26.52
8	READ AND APPROVED THE	75		74	26.52
9	APPROVED THE FINAL MANUSCRIPT	74		73	26.16
10	AS WELL AS THE	72		55	19.71
11	AUTHORS READ AND APPROVED	72		71	25.45
12	IN THE CASE OF	69		49	17.56
13	BEEN SHOWN TO BE	68		51	18.28
14	HAVE BEEN SHOWN TO	67		48	17.20
15	ALL AUTHORS READ AND	66		65	23.30
16	AS SHOWN IN FIGURE	64		34	12.19
17	CLICK HERE FOR FILE	63		32	11.47
18	FOR EACH OF THE	62		31	11.11
19	IS CONSISTENT WITH THE	60		47	16.85
20	IN ADDITION TO THE	57		47	16.85
⋮	⋮	⋮	⋮	⋮	⋮
4,272	MG SAMPLE FROM THE	5		1	0.36
4,273	MM OF EACH DNTP	5		4	1.43

frequency alphabetical statistics filenames notes

4,273 entries Row 35

일반미국영어_공통적다발.lst

File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%
1	IN THE UNITED STATES	125	0.01	64	12.80
2	AT THE SAME TIME	85		68	13.60
3	OF THE UNITED STATES	64		35	7.00
4	THE END OF THE	63		53	10.60
5	FOR THE FIRST TIME	57		51	10.20
6	AT THE END OF	51		44	8.80
7	THE REST OF THE	50		45	9.00
8	IN THE MIDDLE OF	49		45	9.00
9	ONE OF THE MOST	48		43	8.60
10	ON THE OTHER HAND	44		38	7.60
11	AS WELL AS THE	39		37	7.40
12	AT THE UNIVERSITY OF	37		29	5.80
13	IN THE FORM OF	35		33	6.60
14	IN FRONT OF THE	34		29	5.80
15	AS A RESULT OF	33		23	4.60
16	IN THE CASE OF	32		28	5.60
17	IN THE CONTEXT OF	30		21	4.20
18	IS ONE OF THE	29		24	4.80
19	TO THE UNITED STATES	29		22	4.40
20	COMMISSION UNDER THIS CHAPTER	28		1	0.20
⋮	⋮	⋮	⋮	⋮	⋮
1,170	YOU WANT TO SEE	5		4	0.80
1,171	YOUR LOCAL GENERAL SALE	5		1	0.20

frequency alphabetical statistics filenames notes

1,171 entries Row 1,171

그림 2. Biomed 코퍼스와 AmE06코퍼스의 상위 20위와 최하위 2개 어휘다발

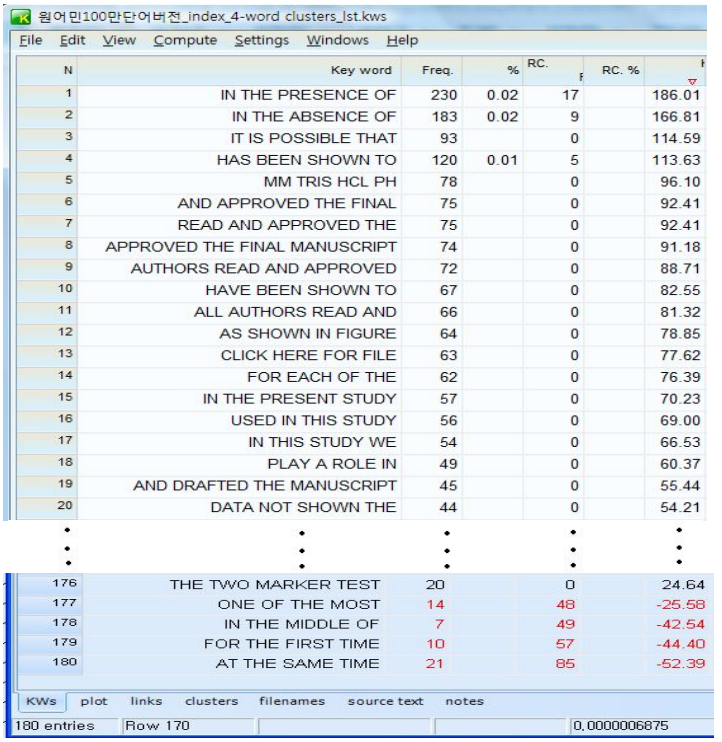
위의 그림 2에서는 공통점과 차이점 둘 다를 찾아 볼 수 있는데 우선 차이점을 논의하자. 상위 20위에 나타난 어휘다발을 살펴보면 두 코퍼스가 다루는 내용이 상당히 다르다는 것을 알 수 있다. Biomed 코퍼스에서는 상위 1위와 2위를 차지하는 *in the presence of* '...존재하에서', *in the absence of* '...부재하에서' 등을 나타내는 과학실험 조건과 관련된 어휘다발이나 3위와 14위에 있는 *have/has been shown to*와 같이 실험 결과와 관련된 표현을 사용하고 있다. 또한

6위와 최하위인 4273위의 *MM TRIS HCL PH, MM of each DNTP* 같은 생물학 관련 전문 용어들이 등장한다. 이에 반하여 참조코퍼스인 AmE06 코퍼스에서 2위, 4위, 5위에 있는 *at the same time, the end of the, for the first time*과 같이 일반영어에서 흔하게 쓰이는 시간표현들과 1위와 3위의 *in the United States, of the United States*와 같은 고유명사와 관련한 표현이 많이 나왔다.

다음으로 두 코퍼스에 나타난 어휘다발의 공통점을 논의하자. 상위 20위 내에서 *on the other hand, as well as the, in the case of* 등 세 개의 어휘다발이 공통으로 발견된 점은 매우 중요한 의의를 갖는다. 이 어휘다발들은 문어체의 담화에 자주 사용되는 것으로 생명공학 분야에서도 이러한 담화표현들을 자주 사용한다는 점이 매우 흥미롭다고 볼 수 있다. 그러나 이러한 어휘다발이 5위, 10위, 12위 등의 상위권에 차지한다고 해서 목표코퍼스만 갖는 문체적인 특징으로 보기에는 상당한 문제가 있다. 그 이유는 참조코퍼스에서도 동일한 어휘다발이 10위, 11위, 16위의 상위권에도 같이 나타나기 때문이다. 이와 같이 어휘다발목록의 상위권에 나타나는 고빈도의 어휘다발만 보아서는 분석하고자 하는 텍스트의 특성을 파악하는 데에는 많은 문제점이 있다는 것을 알 수 있다. 그래서 참조코퍼스와는 다르게 목표코퍼스에서만 많이 나타나거나 너무 적게 나타나는 어휘다발의 특성을 파악하는 것이 중요하다.

그러므로 본 연구의 분석대상인 목표텍스트인 생명공학저널에만 특이하게 나타나는 어휘다발의 목록을 자동으로 추출하여 그러한 어휘다발의 특징을 구조와 기능적인 측면에서 연구할 필요가 있다. 특이한 빈도를 갖는 어휘다발의 추출을 위한 기본 아이디어는 키워드분석과 같은 방법으로 목표코퍼스의 어휘다발의 특성을 연구하기 위해서는 워드스미스 키워드 툴에서 목표코퍼스와 참조코퍼스의 어휘다발목록 간의 대조분석 기능을 갖는 키클러스터목록을 활용하면 성공적으로 원하는 특성을 파악할 수 있다.

통계수식으로 키워드목록을 생성하는 것과 같은 방법으로 위의 그림 2의 각 코퍼스의 어휘다발목록을 이용하여 Biomed 코퍼스의 키클러스터 어휘다발을 추출한 목록은 아래의 그림 3과 같다.


원어민100만단어버전_index_4-word clusters_lst.kws

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	RC.	RC. %	
1	IN THE PRESENCE OF	230	0.02	17		186.01
2	IN THE ABSENCE OF	183	0.02	9		166.81
3	IT IS POSSIBLE THAT	93		0		114.59
4	HAS BEEN SHOWN TO	120	0.01	5		113.63
5	MM TRIS HCL PH	78		0		96.10
6	AND APPROVED THE FINAL	75		0		92.41
7	READ AND APPROVED THE	75		0		92.41
8	APPROVED THE FINAL MANUSCRIPT	74		0		91.18
9	AUTHORS READ AND APPROVED	72		0		88.71
10	HAVE BEEN SHOWN TO	67		0		82.55
11	ALL AUTHORS READ AND	66		0		81.32
12	AS SHOWN IN FIGURE	64		0		78.85
13	CLICK HERE FOR FILE	63		0		77.62
14	FOR EACH OF THE	62		0		76.39
15	IN THE PRESENT STUDY	57		0		70.23
16	USED IN THIS STUDY	56		0		69.00
17	IN THIS STUDY WE	54		0		66.53
18	PLAY A ROLE IN	49		0		60.37
19	AND DRAFTED THE MANUSCRIPT	45		0		55.44
20	DATA NOT SHOWN THE	44		0		54.21
⋮	⋮	⋮		⋮		⋮
176	THE TWO MARKER TEST	20		0		24.64
177	ONE OF THE MOST	14		48		-25.58
178	IN THE MIDDLE OF	7		49		-42.54
179	FOR THE FIRST TIME	10		57		-44.40
180	AT THE SAME TIME	21		85		-52.39

KWs plot links clusters filenames source text notes

180 entries Row 170 0.0000006875

그림 3. AmE06 코퍼스를 참조하여 생성한 Biomed 코퍼스의 키클러스터 어휘다발 목록

위의 그림 3에서 보는 바와 같이 자동으로 생성된 키클러스터 어휘다발은 모두 180개이다. 이 중에 176개는 플러스 핵심도 값을 갖고 단지 네 개의 어휘다발만이 마이너스 핵심도 값을 갖는다. Scott(2012)에 따르면 참조코퍼스와 비교하여 우연확률통계로 예상되는 것보다 더 많이 발생하면 플러스 핵심도 값을 갖는 단어나 어구가 되어 그 값에 아무 표시도 하지 않지만 예상보다 적은 수로 발생하면 그 값에 마이너스를 표시한다. 그러므로 Biomed 코퍼스의 어휘다발의 빈도가 AmE06 코퍼스에 비해서 과도하게 사용(over-use)되면 플러스로, 과소하게 사용(under-use)되면 마이너스로 표시한다.

위의 그림 3의 키클러스터목록에서 발견된 또 다른 흥미로운 점은 어휘다발의 빈도수와 핵심도의 값의 관계는 전혀 관련이 없다는 것이다. 가장 적은 플러스 값인 24.64을 갖는 176위의 *the two marker test* 어휘다발은 출현빈도는 20회이지만 가장 큰 마이너스 값인 -52.39를 갖는 180위의 *at the same time*은 출현빈도가 21회이므로 빈도와 핵심도의 값의 상관관계는 없는 것으로 보인다.

키클러스터목록이 어휘다발목록과 다른 점을 앞서 설명한 예로 들었던 그림 2의 목표코퍼스와 참조코퍼스 둘 다에 상위 20위내에 있었던 *on the other hand, as well as the, in the*

*case of*의 세 개의 어휘다발은 위의 그림 3의 키클러스터목록에는 발견되지 않는다는 것을 확인할 수 있었다.

4.2 키클러스터 어휘다발의 구조적 분석

본 연구에서는 키클러스터목록에 나타난 모든 어휘다발 180개를 분류하였다. 또한 키클러스터 어휘다발이 갖는 특징을 파악하기 위하여 참조코퍼스인 AmE06 코퍼스에서 나타난 상위 180위까지 나타난 어휘다발과도 비교하여 구문 분석하였다. 키클러스터목록에 나타난 동일한 수인 180개의 어휘다발을 분석 기준으로 삼은 이유는 크기에 있다. 목표코퍼스의 특징을 살펴보고자 할 때 참조코퍼스의 크기를 같게 하거나 서로 비슷한 규모로 맞추는 이유와 마찬가지로 한 개의 목표텍스트의 특징을 파악하기 위해서는 참조텍스트와 비교하여 대조분석하려면 텍스트의 크기를 맞추면 어떤 다른 기준화 없이도 목표텍스트의 분석을 가장 효과적으로 할 수 있기 때문이다. 그러므로 Biomed 코퍼스만이 가지고 있는 특성을 보여줄 수 있는 어휘다발의 수가 180개뿐이므로 참조코퍼스인 AmE06의 상위 180개 어휘다발과의 비교를 통해 구조적인 특징과 기능적인 특징을 연구할 수 있다.

아래 표1은 키클러스터목록에 나타난 180개의 모든 어휘다발의 구조적인 특징을 Biber et al.(1999: 996)과 Chen and Baker(2010: 35)를 참고하여 어휘다발 분류방법에 따라 분류한 것이다.

표 1. Biber et al.(1999)와 Chen and Baker(2010)의 분류에 따른 키클러스터 어휘다발의 구조적 분석

Category	Patterns	Biomed		AmE06		Biomed Examples
		type	%	type	%	
NP	(a) noun phrase + post-modifier	58	32%	66	37%	*the presence of a*
PP	(b) preposition + noun phrase	36	20%	76	42%	*in the absence of*
VP	(c) copula be + NP/adjective phrase	6	3%	5	3%	*are present in the*
	(d) VP with active verb	4	2%	0	0	*play a role in*
	(e) anticipatory it + VP/adjective phrase	6	4%	5	3%	*it is likely that*
	(f) passive verb + PP fragment	12	6%	0	0%	*was used as a*
	(g) verb/noun + that-clause fragment	7	5%	1	0.5%	*the possibility that the*
	(h) (verb/adjective) + to-clause fragment	5	3%	3	1.5%	*are likely to be*
	(i) adverbial clause fragment	6	3%	5	3%	*as shown in figure*
	(j) other expressions	40	22%	19	10%	
	total	180	100%	180	100%	

위의 표 1에서 보는 바와 같이 Biber et al.(1999)에서 분류한 방법에서 (i)의 부사절 꼴(adverbial clause fragment)을 추가하여 10가지의 유형으로 분석하였다. Chen and Baker(2010)에 따라 어휘다발의 구조적인 특징은 크게 NP, PP, VP 세 개의 통사범주로 분석하였고 NP와 PP 구조에서는 각 한 개의 유형만 가지나 VP 구조에서는 8개 이상의 세분화된 유형을 갖는다. 키클러스터 어휘다발의 특징은 통사범주의 빈도수를 고려하면 Biomed 코퍼스는 NP-VP-PP의 순으로 빈도 분포를 갖지만 AmE06 코퍼스는 PP-NP-VP의 순서로 분포를 보여준다. 이는 일반영어에서의 어휘다발은 전치사구문이 가장 많이 사용되는 반면에 생명과학저널 영어에서는 명사구문이 가장 많이 사용된다는 것을 반영해준다. 그러나 이러한 결과는 Biber et al.(1999)와 Chen and Baker(2010)에서 보여준 EAP의 고빈도 어휘다발의 통사범주의 빈도분포 PP-NP-VP의 순서는 일반영어 코퍼스와 같지만 Biomed 코퍼스의 키클러스터 어휘다발이 갖는 결과와는 다른 점은 특이하다.

Biomed 코퍼스에서 NP 범주로 나타나는 어휘다발의 예는 *the presence of a, the length of the, a large of number of*와 같은 명사구 + 후치수식어 유형이 58개로 전체의 33%로 가장 많은 분포를 차지한다. PP 범주에서는 *in the presence of, in the absence of, for each of the*와 같은 전치사 + 명사구의 어휘다발이 36개로 20%의 분포를 이루고 있다. VP 범주에서는 다양한 유형을 보여주지만 그 중 가장 두드러지는 유형은 수동동사 + 전치사구 형태의 유형으로 *was added to the, are shown in Figure, was used as a*와 같은 예가 12개로 전체의 6%를 차지하고 있다. 그래서 Biomed 코퍼스에서 나타나는 세 개의 각 통사범주 중 가장 두드러지는 유형인 명사 + 후치수식어의 유형, 전치사 + 명사의 유형, 수동동사 + 전치사구의 유형를 하나씩 아래에서 자세히 설명한다.

4.2.1 명사+후치수식어 유형의 명사구 범주

Biomed 코퍼스와 AmE06 코퍼스에서 가장 높은 비율로 나타난 명사구조의 어휘다발의 실제 사용 양상에 어떠한 차이점이 있는지를 알아보기 위하여서는 아래의 표 2에 나타난 'the + Noun + of the' 구조 안에 사용된 명사의 타입과 토큰의 비율을 살펴보기로 한다.

표 2. 명사구 범주 구조에 나타난 어휘다발

	the + Noun + of + the	type	token	TTR
Biomed	*activity*(24)*, *design*(32), *function*(27), *length*(47), *location*(20), *majority*(39), *presence*(50), *results*(29)	8	268	2.99
AmE06	*use*(11), *top*(19), *size*(12), *side*(14), *rest*(50), *nature*(16), *middle*(26), *history*(17), *front*(13), *end*(63), *edge*(22), *development*(15), *center*(13), *bottom*(19), *beginning*(12), *back*(19)	16	341	4.69

( )*: 괄호 안에 있는 숫자는 해당 명사가 반복적으로 나타난 횟수를 말한다.

위의 표 2를 보면 Biomed에서 타입은 8개이고 토큰은 268개로 나타난 반면에 AmE06에서는 타입 16개, 토큰 341개로 나타났다. 이는 Biomed 코퍼스의 타입에서는 2배, 토큰에서는 1.2배 정도 적게 나타나고 TTR에서도 1.6배 정도 낮은 값을 갖는다. TTR이 시사하는 바와 같이 생명과학저널은 실험 중심의 학술적인 텍스트이므로 어휘가 전문적이고 반복적인 상황이 많아서 한정된 표현이 많이 반복되고 있는 반면에 AmE06 코퍼스인 일반 문어에서는 여러 가지 상황이 존재하기 때문에 보다 다양한 명사가 쓰였다는 것을 알 수 있다.

특히 Biomed 코퍼스에서는 과학적 실험과 관련된 어휘(예: *activity, function, design, results, presence*)와 물리적이고 수량적인 표현(예: *length, location, majority*)과 같은 어휘들이 대부분을 차지하고 있음을 알 수 있다. 이 가운데 아래 예문 (1)과 같이 *the presence of the* '~의 존재 하에서'라는 의미의 *presence*가 50회로 가장 많았다.

(1) **The presence of the** conserved sequence increased the grouping efficiency. (Biomed, 1471-2105-2-9.txt)

반면에 AmE06 코퍼스에서는 일상생활에서 사용되는 단어들로 구성되어 있어 장소나 위치 관련 어휘(예: *top, size, side, middle, front, end, edge, center, bottom, back*)가 주를 이루었다. 가장 많이 사용된 표현으로는 아래 (2)와 같이 *the rest of the* '~의 나머지'라는 의미에서 *rest*가 50회로 가장 많았다.

(2) Working through **the rest of the** room, they examined the catapult Nickford had been working on the day before. (AmE06, P08.txt)

4.2.2 전치사+명사 유형의 전치사구 범주

전치사 범주에서 전치사+명사 유형의 구조는 Biomed 코퍼스에서 36개(20%) AmE06 코퍼스에서 76개(42%)를 사용하고 있어 AmE06이 Biomed보다 2배 이상 많이 사용하고 있다.

하지만 흥미로운 것은 아래 표 3에서 보는 바와 같이 타입은 AmE06이 2배 많이 사용하고 있으나 토큰은 오히려 Biomed가 2배 가량 높다는 사실이다. 즉, 생명공학저널에서는 적은 타입을 사용하되 사용횟수가 매우 높다는 것을 보여준다. 그 이유는 앞선 소절에서 설명한 바와 같이 명사의 경우와 마찬가지로 한정되고 반복된 상황이 자주 등장하기 때문에 고정적이고 동일한 표현을 사용하여 명료하게 전달하고자 하는 텍스트의 특성 때문으로 보인다. TTR에서는 타입의 비율이 2배 높은 일반 문어가 생명공학 텍스트보다 3.8배 높이 나타나서 다양한 상황에서 다양한 유형의 어휘다발이 사용되고 있음을 알 수 있다.

표 3. 전치사구 범주구조에 나타난 어휘다발

in the + Noun + of		type	token	TTR
Biomed	*absence*(183), *amount*(21), *design*(25), [-]**middle*(7), *presence*(230), *regulation*(40)	6	506	1.19
AmE06	*case*(32), *course*(12), *face*(22), *form*(35), *history*(17), *middle*(49), *midst*(24), *number*(12), *presence*(17), *process*(10), *wake*(13), *words*(11)	12	254	4.72

[-]*: 마이너스 핵심도를 갖는 어휘를 지칭한다.

Biomed 코퍼스에서 in the + 명사 + of의 전치사 어휘다발의 유형은 앞서 살펴본 명사구 범주의 구조와 유사하게 실험과 관련된 명사 어휘들이 많이 나타난다. 아래 (3)의 *in the presence of* 전치사 어휘다발 유형에서와 같이 명사 *presence*가 230회로 가장 많이 사용된 어휘이다.

(3) Therefore, neurons were cultured on sections of adult forebrain **in the presence of** dibutyryl-cAMP ... (Biomed, 1471-2202-2-9.txt)

또한 흥미로운 사실은 Biomed에서 마이너스 키워드로 잡힌 *middle*이 있다는 점이다. 구체적으로 *middle*의 빈도를 보면 7회로 선정기준 5회를 조금 넘긴 정도인 반면에 AmE06에서는 *middle*이 49회로 7배 이상 사용되었다. 이렇게 참조코퍼스의 어휘다발이 목표코퍼스보다 특이하게도 너무 많이 사용되기 때문에 마이너스 핵심도의 값이 산출되었다. 아래 (4)는 AmE06의 *in the middle of*이 갖는 용례를 보여준다.

(4) The sheet was blank except for a Web address **in the middle of** the page. (AmE06, N09.txt)

위의 (4)와 같은 AmE06의 전치사구조에서 사용된 명사들은 상황이나 과정 관련의 명사(예: *case, course, process, wake*)가 주를 이루었고 나머지 명사들은 다양한 의미를 표현하므로 한 분류로 묶기는 어려웠다.

4.2.3 수동동사+전치사구 형태의 동사 범주

수동동사 뒤에 전치사구 형태구조를 갖는 동사 범주의 경우 Biomed에서만 12개(6%) 어휘다발이 발견되었지만 AmE06에는 전혀 발견되지 않았다. 이는 생명과학저널에서는 실험의 객관적인 기술이 중요하기 때문에 주체가 드러나지 않는 수동태 문장이 선호되는 반면

에 일반 문어에서는 수동태가 포함된 어휘다발이 빈도 상위에 잡힐 정도로 많이 사용되는 유형이 아님을 알 수 있다.

이러한 수동태구조 중에서 가장 많이 사용된 유형은 아래 (5)와 같이 과거형 수동태인 *was/were performed as described*(103회/109회, 합계 212회)이다.

(5) The GnRH RRA on the membranes **was performed as described** previously. (Biomed, 1471-2121-2-21.txt)

반면에 현재형 수동태에서는 아래 (6)과 같은 *is/are shown in Figure*(91회/88회, 합계 179회)가 가장 많이 사용된 유형이다.

(6) Homologous residues in other LGIC subunits have also been identified and **are shown in Figure** 1. (Biomed, 1471-2091-3-15.txt)

4.3 키클러스터 어휘다발의 기능적 분석

키클러스터 어휘다발의 기능적 분석은 Biber et al.(2004)를 따라서 양태적, 담화조직, 지시적 어휘다발의 세 개의 주요 담화기능 범주로 분류한다. 이는 대학의 강의와 교재를 대상으로 어휘다발을 찾아내어 분류하였으므로 학술문어를 대상으로 하는 본 연구의 어휘다발 분류에도 적합하다는 판단 때문이다.

아래 표 4는 이러한 담화기능 분류기준과 예시를 보여준다.

표 4. Biber et al.(2004)에 따른 어휘다발의 담화기능 분류 기준

Discourse Functions		Biomed		AmE06		Biomed Examples
Stance expressions	Epistemic stance	41	22%	11	6%	*it is possible that*
	Attitudinal/modality stance	3	2%	13	7%	*in order to determine*
Discourse Organizers	Topic elaboration/clarification	7	4%	14	8%	*to be associated with*
Referential Expressions	Identification/focus	5	3%	13	7%	*each of the three*
	Specification of attributes	49	26%	57	31%	*the function of the*
	Time/place/text reference	21	13%	40	22%	*for the first time*
Others		54	30%	32	19%	
TOTAL		180	100	180	100	

위의 표 4는 Biber et al.(2004)의 분류기준을 최대한 단순화한 결과로 담화기능의 세부항목에 전혀 나타나지 않는 것들은 표시하지 않았다. 어휘다발이 갖는 담화기능의 첫 범주인 양태표현은 *it is possible that, it is likely that*와 같은 인식적 양태(41개, 22%)와 *in order to determine, to determine if the*와 같은 의도를 나타내는 태도/양상적 양태(3개, 2%)만 나타났다. 두 번째 범주에 들어가는 담화조직을 나타내는 어휘다발은 *to be associated with, as compared to the*와 같은 내용 부연설명(7개, 4%)은 나타났지만 도입을 나타내는 어휘다발은 나타나지 않았다. 세 번째 담화기능의 범주인 지시적 표현을 나타내는 어휘다발은 *each of the three, to that of that*과 같이 확인을 이끄는 표현(5개, 3%)과 *the function of the, the activity of the*와 같이 명사의 속성에 대한 명시(49개, 26%)가 가장 많이 나타났지만 이러한 담화기능 표현은 AmE06에서도 57개, 31%로 가장 많은 분포를 보여준다. 또한 *for the first time, in this study the*와 같은 어휘다발은 시간과 장소 등을 나타내는 지시적 담화기능(21개, 13%)의 분포를 보여준다. 여기서 흥미롭게도 목표코퍼스와 참조코퍼스를 비교하여 보면 목표코퍼스는 인식적 양태의 담화기능 하나에서만 우위를 점하고 있다.

이를 시각적으로 알아보기 위해서 아래 그림 4와 같은 막대그래프로 비교하여 보았다.

그림 4. 키클러스터 어휘다발 기능분류의 결과

위의 그림 4에서 보는 바와 같이 Biomed 코퍼스에서 AmE06 코퍼스보다 훨씬 두드러지게 많이 나타나는 것은 인식적 양태 기능뿐이다. 다른 담화기능 범주에서는 AmE06 코퍼스가 더 많은 어휘다발이 나타나지만 그 차이가 그렇게 크지 않다.

인식적 양태는 불확실성과 확실성 둘 다를 일컫는 담화기능을 하는데 Biomed 코퍼스와 AmE06 코퍼스의 예를 콘코던스를 통해 찾아보면 아래 (7)과 (8)과 같다.

(7) **It is possible that** p35-C/EBPβ and SRF are not sufficient to form a ternary complex with ... (Biomed, 1471-2121-1-2.txt)

(8) Its proximity to Iran and **the fact that the** majority of its citizens are Shiite Muslims can dominate your vision ... (AmE06, G52.txt)

인식적 양태는 위의 (7)에 나타난 *it is possible that*과 같은 어휘다발은 불확신을 나타나지만 (8)에서와 같이 *the fact that the*와 같은 어휘다발은 확신을 나타낸다. AmE06 코퍼스에서는 (8)과 같은 부류인 *is the fact that, that there is no*와 같은 긍정적이거나 부정적인 확신을 나타내는 인식적 양태 뿐 아니라 구어체 형식인 *I don't know why, I don't know how*와 같은 불확신의 인식적인 양태 기능도 동시에 갖고 있지만, Biomed 코퍼스에서는 (7)의 *it is possible that, it is likely that, may be due to, the possibility that the*와 같은 불확신의 인식적 양태의 담화기능만 가지는 어휘다발을 가지면서 이들이 180개의 키클러스터 어휘다발 목록에서 출현 빈도가 상위 각각 4, 32, 44, 57위에 나타나서 불확신의 인식적 양태가 높은 빈도로 사용된다는 점이 매우 특이하다.[7] 이것은 과학관련 글에서 주로 실험에 근거한 내용을 기술하기 때문에 확신을 나타내는 인식적 양태표현이 주로 사용될 것이라는 예상과는 크게 다르게 나타났다.

5. 결론

지금까지 키클러스터 어휘다발을 추출하기 위하여 워드스미스 프로그램을 사용하여, 목표코퍼스인 Biomed 생명과학저널 코퍼스와 참조코퍼스인 AmE06 현대 미국문어코퍼스의 어휘다발 목록을 각각 생성한 후 Biomed 코퍼스만이 가지는 어휘다발의 특징을 AmE06와 대조 분석하여 구조적으로 그리고 기능적으로 살펴보았다. 본 논문에서는 단어 단위에서 연구하는 키워드분석과 같은 방법으로 어구단위에서는 키클러스터 분석이 특정 텍스트와 코퍼

7) 물론 Biomed 코퍼스에서 긍정적 또는 부정적 확신의 인식적 양태표현을 나타내는 어휘다발도 상당수 (*the fact that the*, 16회; AmE06에서는 30회, *that there is no*, 9회; AmE06에서는 13회) 나타나지만 키클러스터 어휘다발 목록에서는 불확신의 인식적 양태표현만 나타났다는 것이다. 그래서 it is * that을 검색하여 산출된 고빈도 어휘다발 예문 중 상위 10개를 조사한 결과에서도 *에 들어가는 확신의 인식적 양태표현에는 *clear*(23회, 3위), *noteworthy*(13회, 5위), *apparent*(6회, 9위)가 차지하고 불확실의 인식적 양태표현인 *possible*(93회, 1위), *likely*(40회, 2위), *unlikely*(14회, 4위), *conceivable*(12회, 6위), *known*(12회, 6위), *interesting*(9회, 8위), *assumed*(6회, 9위)가 상위 순위에 나타나거나 고빈도를 보여주었다.

스에 나타나는 어휘다발의 특징을 대조 비교 분석하기 위하여 더 적합하고 유용한 새로운 연구방법임을 보여주고자 하였다. 새로운 연구방법인 키클러스터 분석을 이해하기 위하여 용어의 혼동을 초래할 수 있는 키워드, 클러스터와 더불어 키클러스터의 개념에 대해 살펴보았고 어떻게 키클러스터목록을 생성하는가하는 절차와 과정을 자세히 소개하였다.

본 연구에서 분석한 Biomed 코퍼스에 나타난 주요 특성과 키클러스터 어휘다발의 통사범주의 구조적 분석과 담화 기능 분석에서 두드러지게 나타난 주요 특성은 일반언어학에서 예상할 수 있는 것과 그렇지 않은 것을 구분하여 정리하면 다음과 같다. EAP 또는 ESP의 텍스트에서 예상할 수 있는 바와 같이 학술전문저널로서의 생명과학에서 사용된 텍스트도 일반영어의 참조코퍼스와 비교한다면 어휘와 어구의 다양성이 낮을 것이고 어휘다발도 훨씬 정형화된 표현을 많이 사용하고 능동문보다는 수동문을 선호하는 문체의 특성으로 나타날 것으로 예상된다. 이러한 긍정적인 예상결과는 본 연구에서 살펴본 목표코퍼스와 참조코퍼스의 TTR, 어휘다발목록의 수, 어휘다발의 특정 통사 범주내의 명사의 TTR의 비교, 수동구문의 빈도수로 확인할 수 있었다. 첫째, 목표코퍼스인 Biomed 코퍼스에서는 TTR은 2.44, 표준 TTR은 34.41이고 참조코퍼스인 AmE06 코퍼스에서는 TTR이 4.66, 표준 TTR이 45.23으로 나타나 예상된 바와 같이 Biomed 코퍼스가 생명과학저널 전문 학술분야로 어휘와 어구의 다양성이 낮다는 것을 TTR의 비교에서도 확인이 되었다. 둘째, 참조코퍼스에 나타난 어휘다발목록의 총 수가 1,171개인 반면에 Biomed 코퍼스에서는 4,273개나 나타났다. 이러한 결과는 어휘다발 추출 세팅을 5회 이상의 빈도수로 하였으므로 Biomed 코퍼스는 정형화된 표현이 참조코퍼스보다 훨씬 더 많다는 것을 보여주었다. 셋째, 키클러스터 어휘다발의 통사범주 중에서 명사구와 전치사구에 출현되는 명사의 TTR 비교에서도 예상된 결과를 보여주었다. Biomed 코퍼스와 AmE06 코퍼스에서 가장 높은 비율로 나타난 명사구조의 어휘다발의 실제 사용 양상을 'the + Noun + of the' 구조 안에 사용된 명사의 타입과 토큰의 비율을 살펴본 결과 Biomed에서는 TTR이 2.99이고 AmE06에서는 4.69로 위의 전체 코퍼스의 TTR에서 비교한 결과와 같이 Biomed가 한정된 표현을 반복하여 사용한다는 것을 보여주었다. 또한 전치사구 범주 구조에 나타난 어휘다발에서 'in the + Noun + of' 구조 안에 사용된 명사의 타입과 토큰의 비율을 살펴본 결과 Biomed에서는 TTR이 1.19이고 AmE06에서는 4.72로 위와 같은 결과를 보여주었다. 끝으로 수동태구문에 대한 구조적 특징은 Biomed에서 예상된 바이지만 흥미롭게도 AmE06에서는 상위 180개의 어휘다발을 비교한 결과 AmE06에서는 전혀 나타나지 않았다는 점이다. 학술영어에서는 높은 빈도로 수동태구문이 나타나지만 일반영어에서는 빈도 상위에 나타나지 않았다는 점이 특이하였다.

반면에 일반언어학에서 예상되는 바와 다르게 나타난 결과는 어휘다발의 구조분석에서 통사범주의 빈도순서와 기능분석에서는 인식적 양태의 담화기능 등 두 가지 측면에서 나타났다. 첫째, Biber et al.(1999)와 Chen and Baker(2010)에서 보여준 EAP의 고빈도 어휘다발의 통사범주의 빈도분포 PP-NP-VP의 순서는 일반영어의 순서와 같은 결과를 보여주지만,

본 연구에서 다룬 키클러스터 어휘다발의 구조적 특징으로 통사범주의 빈도수를 고려하면 Biomed 코퍼스는 NP-VP-PP의 순의 빈도 분포를 갖지만 AmE06 코퍼스는 PP-NP-VP의 분포를 갖는다. 이는 생명과학저널 이외의 EAP와 일반영어에서의 어휘다발은 전치사구문이 가장 많이 사용되는 반면에 생명과학저널 영어에서는 명사구문이 가장 많이 사용된다는 것이 매우 특이하게 나타났다. 둘째, Biomed 코퍼스에서 키클러스터 어휘다발의 담화 기능을 분석한 결과는 양태적, 담화조직, 지시적 세 개 주요 담화기능 범주로 분류하였을 때 인식적 양태가 가장 두드러지게 나타났다. AmE06 코퍼스에서는 확신과 불확신의 인식적 양태의 두 가지 담화기능을 갖는 어휘다발이 나타났지만 Biomed 코퍼스의 키클러스터 어휘다발목록에서는 *it is possible that, it is likely that, the possibility that the, may be due to*와 같은 불확신의 인식적 양태의 담화기능만 가지는 어휘다발을 가진다는 점이 매우 특이하였다. 이것은 과학관련 글에서 주로 실험에 근거한 내용을 기술하기 때문에 확신을 나타내는 인식적 양태표현이 주로 사용될 것이라는 예상과는 크게 다르게 나타났다.

이상과 같이 생명과학 코퍼스에 나타난 주요 특성과 키클러스터 어휘다발의 통사범주의 구조적 분석과 담화 기능 분석에서 일반 언어학적으로 예상되거나 특이한 특징을 고려해보면 본 연구에서 새로운 연구방법으로 도입한 키클러스터 분석이 특정 텍스트와 코퍼스에 나타나는 어휘다발의 특징을 대조 비교 분석하기에는 아주 유용한 새로운 연구방법임을 보여준다.

참고문헌

김혜영 · 강범모. (2011). 신문 사설의 어휘적 특징: 2009년 신문 코퍼스에 기초한 키워드 연구. *담화인지언어학회, 18*(3), 89-113.

장세은 · 변현정. (2011). 코퍼스를 활용한 해사영어 어휘분석. *새한영어영문학, 53*(4), 247-268.

전지은. (2010). 성별에 따른 한국어 부사 사용 양상: 세종 구어 말뭉치를 활용하여. *언어와 언어학, 47*, 191-218.

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes, 31*, 81-92.

Baker, P. (2004). Querying keywords: Questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics, 32(4)*, 346-359

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14(2)*, 30-49.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*, 393-423.

Culpeper, J. (2002). Computers, language and characterisation: An analysis of six characters in Romeo and Juliet. In U. Melander-Marttala, C. Ostman, & M. Kyto (Eds.), *Conversation in Life and in Literature: Papers from the Association Swedoise de Linguistique Appliquee(ASLA) Symposium, 15* (pp.11-30). Universitetstryckeriet, Uppsala, Sweden.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics, 14(1)*, 29-59.

Jhang, S., & Hong, S. (2012). A corpus-based analysis of English abstracts of Korean scholarly articles: A case study of English literature and English linguistics. *The New Korean Journal of English Language & Literature, 54(4)*, 289-306.

Jhang, S., & Parent, K. (2011). The vocabulary of Maritime English: Keyword analysis of the English homepages of port authorities around the world. *Korean Journal of English Language and Linguistics, 11(4)*. 1065-1083

Johnson, S., Culpeper, J. & Suhr, S. (2003). From politically correct councillors to Blairte nonsense: Discourses of political correctness in three British newspapers. *Discourse and Society, 14(1)*, 29-47.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4-21.

Leech, G., & Fallon, R. (1992). Computer corpora: What do they tell us about culture? *ICAME Journal, 16*, 29-50.

Mahlberg, M. (2007). Clusters, key clusters and local textual functions in Dickens. *Corpora, 2(1)*, 1-31.

Meunier, F., & Granger, S. (2008). *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins.

Oakes, M. P., & Farrow, M. (2007). Use of the chi-squared test to examine

vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistics Computing, 22(1)*, 85-99.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13(4)*, 519-549

Rayson, P., Leech, G., & Hodges, M. (1997). Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics, 2(1)*, 133-152

Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. In K. Ajimer & B. Altenberg (Eds.), *English Corpus Linguistics* (pp. 128-143). Cambridge: Cambridge University Press.

Rizzo, C. R. (2010). Getting on with corpus compilation: From theory to practice. *ESP World, 1(9)*, 1-23.

Sardinha, T. B., & Shimazumi, M. (2003). Schoolchildren writing: A corpus-based analysis. *Linguagem & Ensino, 6(1)*, 11-33.

Scott, M. (1999). *WordSmith tools user help file*. Oxford: Oxford University Press.

Scott, M. (2005). The behavior of key words. *Corpus Linguistics Birmingham*, 1-8.

Scott, M. (2012). *WordSmith tools help*. Liverpool: Lexical Analysis Software.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Sinclair, J. (1991). *Corpus. concordance, collocation*. Oxford: Oxford University Press.

Starcke, B. (2009). Keywords and frequent phrases of Jane Austen's Pride and Prejudice. *International Journal of Corpus Linguistics, 14(4)*, 492-523.

Tribble, C. (2000) Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the third international conference on teaching and language corpora* (pp. 75-90). Hamburg: Peter Lang.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics, 21(4)*, 463-89.

장세은
606-791 부산광역시 영도구 동삼동 1번지
한국해양대학교 국제대학 영어영문학과
전화: (051) 410-4595
이메일: jhang@hhu.ac.kr

이성민
608-737 부산광역시 남구 용소로 45 (608-737)
부경대학교 인문대학 영어영문학과
전화: (051) 629-5370
이메일: roy7942@hanmail.net

Received on October 25, 2012
Revised version received on November 28, 2012
Accepted on November 28, 2012