Journal of Digital Convergence
Vol. 19. No. 7, pp. 33-39, 2021

ISSN 2713-6434 / eISSN 2713-6442
https://doi.org/10.14400/JDC.2021.19.7.033

키워드 분석에 대한 최신 접근법 비교 연구: 성경 코퍼스를 중심으로

하명호
신라대학교 교양과정대학 교수

A Comparative Study of a New Approach to Keyword Analysis: Focusing on NBC

Myoungho Ha
Assistant Professor, College of General Education, Silla University

요 약 본 연구는 구약 성경 코퍼스와 신약 성경 코퍼스, 그리고 구약과 신약 성경을 통합한 코퍼스에서 추출된 키워드 목록의 어휘적 특징을 분석하고, 또 사용빈도 기반의 키워드 분석보다 분포도 기반 키워드 분석이 더 우수한 분석 방식임을 밝히고자 하였다. 이를 위해 Bible Hub의 NLT 웹사이트에서 성경 파일을 다운받아 약 57만 어절의 구약 성경 코퍼스와 약 20만 어절의 신약 성경 코퍼스를 구축하였다. 목표 코퍼스와 참조 코퍼스의 비교를 통한 키워드 목록을 추출하기 위해서 Scott(2020)의 *WordSmith 8.0* 프로그램을 사용하였다. 그 결과, 분포도 기반 키워드 분석이 사용빈도 기반의 키워드 분석보다 키워드 목록의 어휘적 특징을 보다 더 잘 나타낼 수 있었고, 또 코퍼스 내용의 대표성과 변별성을 충분히 충족시킬 수 있는 최적의 키워드 목록을 추출하기 위해서는 분포도 기반 키워드 분석이 더 우수한 방식임을 밝혔다.

주제어 : 키워드 분석, 사용빈도, 분포도, 어휘적 특징, 내용의 대표성, 내용의 변별성

Abstract This paper aims to analyze lexical properties of keyword lists extracted from NLT Old Testament Corpus(NOTC), NLT New Testament Corpus(NNTC), and The NLT Bible Corpus(NBC) and identify that text dispersion keyness is more effective than corpus frequency keyness. For this purpose, NOTC including around 570,000 running words and NNTC about 200,000 were compiled after downloading the files from NLT website of Bible Hub. Scott's (2020) *WordSmith 8.0* was utilized to extract keyword lists through comparing a target corpus and a reference corpus. The result demonstrated that text dispersion keyness showed lexical properties of keyword lists better than corpus frequency keyness and that the former was a superior measure for generating optimal keyword lists to fully meet content-generalizability and content distinctiveness.

Key Words : keyword analysis, frequency, dispersion, lexical properties, content-generalizability, content-distinctiveness

*Corresponding Author : Myoungho Ha(hadash21@silla.ac.kr)
Received April 26, 2021 Revised May 11, 2021
Accepted July 20, 2021 Published July 28, 2021

1. 서론

EFL(English as a Foreign Language)이나 ESL(English as a Second Language) 학습을 통해서 비원어민 학습자들이 최종적으로 습득하려는 것이 의사소통능력(communicative competence)의 향상이라고 한다면, 이 능력을 배양하는데 있어서 가장 기본이 되는 것은 당연히 풍부한 어휘 지식이라고 할 수 있다. 이렇듯 언어 습득에 있어서 어휘 지식은 이미 많은 연구들에서 그 중요성이 논의되었다[1-6]. 특히 Nation(1990, 1994)에 따르면 학습자는 어휘 지식을 통해서 어휘의 형태나 위치, 그리고 기능 및 의미를 이해할 수 있고, 또한 언어의 4가지 능력 – 듣기, 말하기, 읽기, 쓰기 – 을 실제 언어사용 환경에서 효과적으로 활용할 수 있게 한다고 하였다[7].

코퍼스 언어학에서 널리 통용되는 키워드(keyword)는 참조 코퍼스(reference corpus)와 비교해서 목표 코퍼스(target corpus)에서 현저한 사용빈도(frequency)를 보이는 어휘들을 가리키는데, Baker(2004)에 의하면 키워드는 비교 대상 코퍼스들 간의 내용적 특징(aboutness)이나 문체(style)에 있어서 중요한 어휘적 차이를 나타낸다고 하였다[8]. 아울러 키워드는 담화나 텍스트 그리고 장르 분석 등에서 필수적인 도구인 키워드 분석을 통해서 추출되는데, 키워드 분석은 코퍼스 언어학에서 가장 널리 활용되는 분석 방법들 중 하나로 특정 담화 영역(specific discourse domain)에서 핵심적인 어휘들을 밝혀내는 간단한 통계적 방법을 가리킨다[9, 10].

이렇듯 키워드는 비교 분석하려는 텍스트나 코퍼스들 간의 내용적 특징이나 문체상의 두드러진 차이를 밝히는 핵심적인 어휘로써 코퍼스 언어학에서 중요한 역할을 한다고 볼 수 있다. 지금까지 키워드는 비교 대상 코퍼스에서 통계적으로 더 높은 사용빈도를 보이는 어휘들을 추출하였다. 하지만 Egbert & Biber(2019)에 따르면 이렇게 어휘들의 사용빈도에만 기반을 두고 추출한 키워드들이 해당 텍스트나 코퍼스들의 담화 영역을 대표하지 못하는 한계점이 있었다. 따라서 본 연구에서는 이런 문제점을 해결하기 위하여 Egbert & Biber(2019)의 주장에 따라 기존의 사용빈도에 기반을 둔 키워드 분석보다는 어휘가 해당 텍스트나 코퍼스에서의 분포 정도를 나타내는 분포도(dispersion)에 기반을 두고 키워드 분석을 하고자 한다.

지금까지 교육, 공학, 생물학, 농학, 항공, 해양, 비즈니스, 금융, 의학 등의 다양한 전문 학술분야에 대한 키워드의 추출 및 분석은 이루어졌지만 구약과 신약 성경으로 구축된 코퍼스를 기반으로 이루어진 연구가 거의 없었다[11-14]. 예를 들어 영어 성경기반 코퍼스 분석을 살펴보면 어휘의 연어 관계에 대한 연구가 거의 유일하다. 유문영(2019)의 'NIV 영어성경에서 사용된 *Make* 의미와 연어 연구'에서는 NIV 영어성경 4복음서에 사용된 *make* 동사의 사용빈도수, 의미 및 연어들을 분석하고 있다[15]. 따라서 본 연구에서 분포도 기반의 최신 키워드 분석 방식을 활용하여 구약과 신약 성경 코퍼스에서 코퍼스 내용의 대표성(content-generalizability)과 코퍼스 내용의 변별성(content-distinctiveness)을 갖춘 최적의 키워드를 추출하려는 연구는 충분한 연구의 가치가 있을 것으로 사료된다.

본 연구의 연구 질문을 살펴보면 다음과 같다. 첫째, 구약 성경 코퍼스(NLT[1] Old Testament Corpus, NOTC), 신약 성경 코퍼스(NLT New Testament Corpus, NNTC) 및 구약과 신약 성경을 통합한 코퍼스(The NLT Bible Corpus, NBC)의 키워드 분석을 통해 어휘적 특징을 밝힌다. 둘째, 코퍼스의 분포도 기반 키워드 분석이 최적의 키워드 목록을 추출하는데 있어서 기존의 전통적인 사용빈도 기반의 키워드 분석보다 더 우수한 분석 방식임을 밝힌다.

2. 이론적 배경

2.1 사용빈도 기반 키워드 분석

사용빈도에 기초한 전통적인 키워드 분석을 살펴보면 Scott(2020)의 *WordSmith 8.0* 프로그램이 개발되기 이전의 버전들(*WordSmith 5.0, 6.0, 7.0*)에서 주로 이루어졌다[16]. 다시 말해서, 이전의 버전들은 Dunning(1993)의 log-likelihood(LL)에 의해 계산된 키워드가 핵심도(keyness)값을 가진다고 간주하고, 유의확률(p-value) 0.05 이하에서 LL의 계산 결과 값인 핵심도 값이 3.84 이상이 되는 어휘들만 키워드로 선택하였다[17]. 이렇게 선택된 키워드 목록을 통해 분석 대상 텍스트나 코퍼스의 특징을 원어민의 직관(intuition)보다

1) 현재 영어성경 번역은 50개 이상이 사용되고 있는데, 그 중에서 NLT(The New Living Translation Bible)는 KJV(King James Version), NIV(New International Version), RSV(Revised Standard Version)와 함께 가장 많이 읽히는 번역이다.

더 객관적인 방식으로 분석할 수 있었다. 특히 키워드 목록이 참조 코퍼스와 비교해서 목표 코퍼스에서 더 높은 사용빈도를 보일 경우 긍정의 핵심도 값을 가지므로 긍정의 키워드로 분류하고, 그 반대의 경우는 부정의 핵심도 값을 가지므로 부정의 키워드로 분류하였다[18].

하지만 상기에서 언급한 유의확률에 기반을 둔 분석방식은 추출된 키워드들의 중요도를 적절히 평가할 수 없는 문제점이 있었다. 예를 들어 목표 코퍼스에서 중요한 키워드로 분류되었지만 참조 코퍼스에서 높은 사용빈도를 보일 경우 핵심도 값이 낮게 책정되고, 오히려 그 반대의 경우는 핵심도 값이 높게 나오므로 핵심도 값이 서로 다른 키워드들의 중요도 차이를 평가하는 객관적인 기준이 될 수 없었다. 즉, 높은 핵심도 값이 목표 코퍼스와 더 강력하게 연관되어 있다는 가정에 근거하여 핵심도 값에 따라 키워드의 순위를 결정하게 되면 해당 코퍼스에서의 키워드의 중요도를 적절히 반영하지 못하는 문제점이 있을 수 있다. 따라서 Scott(2020)의 *WordSmith 8.0* 프로그램에서는 LL 통계의 대안으로 BIC score를 사용하여 유의확률 방식으로 추출하는 키워드의 한계점을 보완한 알고리즘을 제시하였다.

2.2 분포도 기반 키워드 분석

상기에서 언급한 사용빈도 기반 키워드 분석의 문제점을 해결하기 위해 Egbert & Biber(2019)는 텍스트의 분포도[2]에 기반을 둔 새로운 분석 방법을 활용하여 키워드를 추출하였다. 특히 코퍼스의 키워드를 평가하는 2가지 기준으로 코퍼스 내용의 대표성과 변별성을 제시하였고, Baker(2004, 2010) 및 Culpeper(2009)에서 논의된 사례들은 코퍼스 내용의 대표성 혹은 변별성이 결여되었음을 잘 보여준다고 하였다[19, 20].

Baker(2004, 2010)에서 코퍼스를 구성하는 여러 텍스트들 중에서 하나의 텍스트에서 출현해도 사용빈도만 높으면 키워드로 선택되는 사례가 있었고, 또 키워드로 선택된 어휘들의 수가 너무 적어도 참조 코퍼스와 비교해서 코퍼스 내용의 변별성이 확보되는 사례가 있었는데, 2가지 사례 모두 목표 코퍼스의 내용의 대표성이 결여되었다고 분석하였다. Culpeper(2009)에서는 셰익스피어의 로미오와 줄리엣에 등장하는 여섯 명의 등장인물들 각각에 대해 사용빈도 기반의 키워드 목록을 추출하였다. 하지만 *a, of, the, an, that*과 같은 기능어(function/grammatical words)[3]들이 상위 빈도를 차지하였는데, 이런 기능어들은 일반적인 고빈도 기능어들이므로 개별 등장인물들이 발화한 담화의 내용적 특징들과 연관시키는 것이 어려워 코퍼스 내용의 변별성이 결여되었다고 분석하였다.

이렇듯 사용빈도를 완전히 배제하고 분포도 기반의 새로운 키워드 분석 방식을 활용하면 기존의 사용빈도 기반의 키워드 분석에서 잘못 선택된 키워드들을 적절히 걸러낼 수 있고, 또한 더 효과적인 방식으로 최적의 키워드를 추출할 수 있다는 점에서 기존의 키워드 분석 방식보다 더 뛰어난 방식이라고 주장하였다.

3. 연구 방법

본 연구에서는 키워드 분석에 대한 최신 접근법을 비교하기 위해 NLT 구약과 신약 성경으로 구성된 코퍼스를 구축하였다. 구약 성경 코퍼스인 NOTC는 율법서, 역사서, 시가서, 대선지서 및 소선지서로 구성되어 있고, 신약 성경 코퍼스인 NNTC는 복음서, 바울서신, 공동서신 및 역사서/예언서로 구성되어 있다. 특히 NOTC와 NNTC를 구축하기 위해 본 연구는 Bible Hub의 NLT 웹사이트(https://biblehub.com/nlt/genesis/4.htm)에서 성경 파일을 다운받은 후, nPDF 프로그램을 사용하여 텍스트 파일로 변경하였고, 또 코퍼스 구축에 불필요한 부분들을 정리하기 위해 정규식(R-expressions)을 사용하였다.

한편 키워드를 추출하기 위해 NOTC와 NNTC를 목표 코퍼스로 활용하였고, 참조 코퍼스는 총 4백만 어절의 BNC Baby를 구성하고 있는 4개 분야 – 구어(spoken), 학술(academic), 소설(fiction), 뉴스(newspapers) – 중에서 학술분야(Aca)를 활용하였다. 목표 코퍼스와 참조 코퍼스에 대한 구체적인 통계 정보는 다음 Table 1에서 보듯이 Scott(2020)의 *WordSmith 8.0* 프로그램을 사용하여 추출하였다.

Table 1. Statistical information of target and reference corpus

Corpus		Tokens	Types	TTR	STTR
Target	NOTC	569,109	12,230	2.15	34.15
	NNTC	191,292	6,524	3.41	34.95
Reference	BNC Baby (Aca)	1,014,027	33,532	3.37	40.21

2) Gries(2021: 4)에 의하면 Egbert & Biber(2019)에서 언급한 분포도는 결국 사용범주(range)를 의미한다고 하였다.

3) 기능어는 조동사, 전치사, 접속사, 한정사, be동사, 대명사로 구성되어 있고, 내용어(content words)는 명사, 동사, 형용사, 부사, 부정어, 감탄사, 의문사 등으로 구성되어 있다.

Table 1에서 NOTC의 크기는 NNTC에 비해 거의 3배가량 크지만 해당 코퍼스에서의 어휘의 다양성을 나타내는 수치인 어휘밀도(TTR)를 살펴보면 반대의 결과를 보이고 있다. 하지만 비교 분석하고자 하는 NOTC와 NNTC의 크기가 다를 경우에는 TTR보다 표준 어휘밀도(STTR)를 비교하는 것이 더 정확하다. 구체적으로 NOTC와 NNTC의 TTR은 2.15와 3.41로 큰 차이를 보이지만 STTR의 경우 34.15와 34.95로 그 차이가 크지 않음을 알 수 있다. 이 결과는 NOTC보다 NNTC의 어휘의 다양성이 더 높다는 표시이므로 구약 성경보다 신약 성경에서 더 풍부한 어휘가 사용되었음을 알 수 있다. 한편 일반영어로 구성된 BNC Baby (Aca)의 STTR이 40.21로 상당히 높은 것은 구약과 신약 성경이라는 특정 분야의 영어로 구성된 NOTC와 NNTC보다 훨씬 더 다양한 어휘들이 사용되었음을 의미한다.

본 연구에서는 키워드 분석을 위해 기존에 사용하던 Scott(2016)의 *WordSmith 6.0* 프로그램 대신에 키워드 추출에 있어서 새로운 방식을 활용한 Scott(2020)의 *WordSmith 8.0* 프로그램을 사용하였다[21]. 즉, 기존의 키워드 분석은 단순히 어휘의 사용빈도에 기반을 둔 접근법이라면 Scott(2020)의 *WordSmith 8.0* 프로그램에서는 사용빈도뿐만 아니라 사용범주에 해당하는 텍스트의 분포도를 활용할 수 있어 더 적절한 키워드 분석을 할 수 있기 때문이다.

4. 분석 결과 및 논의

4.1 NOTC와 NNTC의 사용빈도 기반 키워드 분석

기존의 사용빈도 기반의 분석 방식을 활용하여 NOTC와 NNTC에서 키워드를 추출하였고, 다음 Table 2는 상위 20위에 해당하는 키워드 목록을 보여준다.

사용빈도 기반 키워드 분석으로 추출된 키워드를 살펴보면 NOTC와 NNTC에서 공통으로 사용되는 어휘는 총 20개 중에서 14개인 반면 서로 다른 어휘는 볼드체의 이탤릭체로 표시된 6개에 불과하다. 더구나 14개의 공통 어휘들 중에서 3개의 내용어(*God, Lord, son*)를 뺀 나머지 11개는 일반영어에서도 흔히 사용되는 고빈도 기능어들이다. 물론 기능어들 중에서 일부는 해당 코퍼스의 문법적 패턴 및 특징을 분석하는데 기여하기도 한다.

하지만 코퍼스 내용의 대표성 및 변별성을 충족시키기 위해서는 내용어가 충분히 추출되어야 하는데, 기능어가 이렇게 많다는 것은 특수 코퍼스인 성경 코퍼스와 일반 영어로 구성된 BNC Baby (Aca) 사이뿐만 아니라 NOTC와 NNTC 사이의 내용의 대표성 및 변별성을 충분히 충족시키지 못하는 문제점이 있다.

Table 2. Keywords top-ranked 20 by frequency of NOTC and NNTC

	NOTC		NNTC	
N	Keywords	Texts	Keywords	Texts
1	Lord	36	you	27
2	you	39	God	27
3	will	39	***Jesus***	26
4	your	39	I	27
5	I	39	him	26
6	my	39	he	27
7	God	36	me	22
8	me	37	***who***	26
9	***Israel***	35	will	27
10	he	37	your	27
11	***king***	35	***Christ***	23
12	them	39	Lord	24
13	***people***	38	they	26
14	him	36	***don***	25
15	son	34	***said***	18
16	his	38	my	25
17	they	39	them	26
18	***land***	38	his	25
19	***all***	39	son	23
20	***David***	19	***father***	26

다음으로 NOTC와 NNTC에서 추출된 내용어의 어휘적 특징을 살펴보면 구약은 이스라엘의 역사서 역할을 하므로 *Israel*이 아주 빈번히 등장하고, 또 두 번째 왕(*king*)이면서 구약 성경을 통틀어 가장 유명한 다윗왕(*David*)이 구약에 많이 등장하므로 높은 사용빈도를 보이고 있다. 신약은 구세주(*Christ*) 예수님(*Jesus*)과 하나님 아버지를 칭하는 *God, father* 및 *Lord*가 높은 사용빈도를 보인다. 하지만 *God*과 *Lord*는 NOTC와 NNTC에 모두 출현하므로 변별성이 결여되어 있고, 또 키워드로 추출된 어휘들이 너무 적을 경우 앞에서도 언급했듯이 대표성이 결여될 수 있다. 이처럼 사용빈도 기반의 키워드 분석으로 추출된 키워드 목록을 통해서 어휘적 특징을 어느 정도 살펴볼 수는 있지만 이들 소수의 어휘들이 NOTC와 NNTC의 내용을 충분히 대표할 수 있을지 의문시 된다.

4.2 NOTC와 NNTC의 분포도 기반 키워드 분석

Table 3에서는 분포도에 기반을 둔 새로운 키워드 분

석 방식을 활용하여 키워드를 추출하였고, 상위 20위에 해당하는 키워드 목록을 보여준다.

Table 3. Keywords top-ranked 20 by dispersion of NOTC and NNTC

	NOTC		NNTC	
N	Keywords	Texts	Keywords	Texts
1	***Judah***	33	***Jesus***	26
2	sin	28	don	25
3	don	34	***pray***	20
4	***armies***	27	sin	20
5	***Gods***	27	***greetings***	20
6	***wilderness***	26	holy	23
7	***sinned***	25	***believers***	18
8	***covenant***	25	***joy***	22
9	holy	31	***righteous***	17
10	***Israel***	35	***grace***	21
11	didn	24	***righteousness***	15
12	***heavens***	24	***glory***	22
13	***goats***	24	***doesn***	14
14	***flocks***	24	didn	14
15	***swords***	24	***amen***	14
16	***Moab***	23	***hearts***	21
17	***pour***	23	***sins***	21
18	honor	29	honor	18
19	***vineyards***	22	***Christ***	23
20	***Edom***	22	***sinful***	13

분포도 기반 키워드 분석의 두드러진 특징은 Table 3에서 보듯이 기능어가 전혀 없고, 모두 내용어로 이루어져 있다는 점과 공통 어휘는 5개 – *don, sin, holy, didn, honor* – 뿐인 반면 서로 다른 어휘가 15개에 이른다는 점이다. 이것은 NOTC와 NNTC의 내용의 대표성뿐만 아니라 변별성을 충분히 충족시킬 수 있다는 점에서 사용빈도 기반의 키워드 분석과 비교해서 현저한 차이를 보여준다고 할 수 있다.

구체적으로 NOTC와 NNTC의 어휘적 특징을 살펴보면 다음과 같다. 이스라엘(*Israel*)의 역사서인 구약에서는 다양한 지역들(*Judah, Moab, Edom*)에서 잡다한 신들(*Gods*)을 섬기는 이야기와 전쟁이야기(*swords, armies*)가 많이 나오고, 또 구약시대에 재산의 일부이자 제물로 바치기도 하였던 동물들(*goats, flocks*)이 빈번히 등장하고 있다.

반면에 4권의 복음서를 통해서 하나님 아버지의 말씀을 구세주(*Christ*) 예수님(*Jesus*)이 인간에게 전하는 내용과 사도들이 사람들에게 예수님의 말씀을 전하는 내용으로 구성되어 있는 신약에서 하나님의 특성을 나타내는 표현들(*righteous, grace, righteousness*)과 죄를 지은 사람들이 기도를 통해 구원받고 하나님에게 감사를 드리는 표현들(*pray, greetings, believers, joy, glory, amen, sins, sinful*)이 높은 사용빈도를 보이고 있다.

이렇게 NOTC와 NNTC에서 각각 추출된 내용어들을 분석하면 구약과 신약을 대표하는 어휘들의 대표성뿐만 아니라 구약과 신약을 구별 짓는 어휘들의 변별성도 살펴볼 수 있다. 따라서 Egbert & Biber(2019)의 주장처럼 각 코퍼스에서 추출된 내용어들의 어휘적 특징을 통해서 각 코퍼스의 내용의 대표성과 변별성을 나타낼 수 있다는 점에서 분포도 기반 키워드 분석 방식이 키워드 기반 분석 방식보다 더 뛰어난 방식이라고 사료된다.

4.3 NBC의 사용빈도와 분포도 기반 키워드 비교 분석

다음은 구약과 신약 성경을 통합한 코퍼스인 NBC에서 사용빈도 기반 분석 방식과 분포도 기반 분석 방식을 활용하여 키워드를 추출하고, 그 차이점을 비교 분석하고자 한다. NBC에 대한 사용빈도 기반 분석 방식으로 추출된 총 3,442개의 키워드 목록들 중에서 긍정의 키워드

Table 4. Positive keywords top-ranked 20 by frequency and dispersion of NBC

	Corpus frequency keyness		Text dispersion keyness	
N	Keywords	Texts	Keywords	Texts
1	you	66	don	59
2	Lord	60	sin	48
3	will	66	holy	54
4	I	66	didn	38
5	God	63	joy	47
6	your	66	honor	47
7	he	64	Judah	37
8	my	64	pray	37
9	me	59	righteous	35
10	him	62	covenant	35
11	***Israel***	49	heavens	34
12	king	47	wilderness	34
13	them	65	sinned	34
14	they	65	sins	50
15	people	65	prophet	33
16	his	63	blessed	33
17	son	57	***Israel***	49
18	Jesus	26	armies	32
19	who	64	doesn	32
20	said	53	Gods	31

목록은 1,617개이고, 부정의 키워드 목록은 1,825개이다. 분포도 기반 분석 방식으로 추출된 총 2,056개의 키워드 목록들 중에서 긍정의 키워드 목록은 524개이고, 부정의 키워드 목록은 1,532개이다. 다음 Table 4는 긍정의 키워드 목록들 중에서 상위 20위 키워드 목록을 보여주고 있다.

긍정의 키워드 목록이 참조 코퍼스인 BNC Baby (Aca)와 비교해서 목표 코퍼스인 NBC에서 가장 빈번하게 사용되는 어휘를 가리킨다는 것을 고려했을 때, Table 4에서 가장 두드러진 것은 사용빈도와 분포도 기반 상위 20위 키워드 목록들 중에서 공통 어휘는 *Israel*뿐이라는 점이다. 이것은 2가지 키워드 분석 방식 간의 분명한 차이를 확인할 수 있게 한다는 점에서 의의가 있다.

구체적으로 사용빈도 기반 분석 방식으로 추출된 어휘들을 살펴보면 총 20개 중에서 기능어는 12개 - *you, will, I, your, he, my, me, him, them, they, his, who* - 이고, 나머지 8개는 내용어로 구성되어 있다. 이것은 4.1에서도 언급했듯이 기능어가 너무 많으면 분석 대상인 코퍼스 사이의 내용의 대표성 및 변별성이 결여되는 문제점이 있다. 반면에 분포도 기반의 분석 방식으로 추출된 어휘들을 보면 기능어는 전혀 없고 모두 내용어로 구성되어 있는데, 이것은 일반영어로 구축된 BNC Baby (Aca)에서는 드물지만 성경으로 구축된 특수 코퍼스에서 흔히 접할 수 있는 어휘들이라는 점에서 성경 코퍼스만의 어휘적 특징을 엿볼 수 있다.

다시 말해서 Table 2와 3에서 언급했듯이 분포도 기반의 분석 방식으로 추출된 어휘들을 살펴보면 구약과 신약 성경의 내용을 각각 대표하면서 일반영어로 구축된 참조 코퍼스와의 차이를 보여주는 변별성도 보여준다는 점에서 분포도 기반 키워드 분석 방식의 우수성을 확인할 수 있다.

Table 5에 제시된 부정 키워드 목록은 참조 코퍼스인 BNC Baby (Aca)와 비교해서 NBC에서 거의 사용되지 않는 어휘들을 가리키는데, 사용빈도와 분포도 기반의 2가지 분석 방식에서 공통적으로 사용된 어휘는 2개 – *particular, general* – 이고, 나머지 18개는 서로 다른 어휘들이 추출되었음을 알 수 있다.

구체적으로 사용빈도 기반 분석 방식으로 추출된 어휘들은 총 12개 -*however, between, more, or, that, by, an, of, a, in, is, which* – 가 기능어이고, 8개가 내용어로 구성되어 있는 반면에 분포도 기반 분석 방식으로 추출된 어휘들은 모두 내용어로 구성되어 있어 코퍼스 내용의 대표성과 변별성 측면에서 2개의 분석 방식

Table 5. Negative keywords top-ranked 20 by frequency and dispersion of NBC

	Corpus frequency keyness		Text dispersion keyness	
N	Keywords	Texts	Keywords	Texts
1	however	27	described	4
2	history	14	developed	4
3	form	16	probably	4
4	between	41	useful	5
5	***particular***	3	basic	3
6	different	24	independent	3
7	example	31	complex	3
8	***general***	5	familiar	3
9	more	60	basis	3
10	or	62	widely	3
11	such	54	reasons	3
12	that	66	structure	4
13	by	64	significance	4
14	an	62	***general***	5
15	of	66	introduced	3
16	a	66	difficulty	3
17	in	66	attempt	3
18	is	66	similarly	3
19	which	55	system	3
20	#	34	***particular***	3

사이의 분명한 차이를 확인할 수 있다. 특히 Table 5에 표시된 내용어들이 일반영어에서 쉽게 접할 수 있는 어휘들이지만 NBC에서는 사용빈도가 가장 낮은 어휘들이라는 점에서 NBC가 일반영어가 아닌 성경과 밀접히 연관된 특수 코퍼스임을 잘 보여준다고 할 수 있다.

이상에서 사용빈도 및 분포도 기반 분석 방식을 활용하여 추출된 긍정 및 부정의 키워드 목록을 살펴보았다. 전자의 분석 방식으로 추출된 긍정과 부정의 키워드 목록은 많은 기능어들을 포함하고 있어서 코퍼스 내용의 대표성과 변별성 결여를 초래할 수 있는 문제점이 있었다. 따라서 본 연구에서는 Egbert & Biber(2019)가 제안한 분포도 기반 키워드 분석 방식이 일반영어에서 흔히 접할 수 있는 기능어보다는 NBC의 어휘적 특징을 더 잘 밝힐 수 있는 내용어들을 추출한다는 점에서 사용빈도 기반 키워드 분석 방식보다 더 우월하다는 것을 밝히고자 한다.

5. 결론

본 연구는 사용빈도 및 분포도 기반의 키워드 분석 방

식을 활용하여 NOTC와 NNTC 및 NBC에서 각각 추출된 키워드 목록의 어휘적 특징을 분석하였는데, 일반영어에서도 많이 접할 수 있는 고빈도의 기능어가 많이 포함된 전자의 분석 방식은 코퍼스 내용의 대표성과 변별성을 충분히 충족시키는데 문제점이 있었다. 따라서 코퍼스 내용의 대표성과 변별성을 갖춘 최적의 키워드 목록을 추출하기 위해서는 Egbert & Biber(2019)가 제안한 분포도 기반 키워드 분석 방식이 기존의 키워드 기반 분석 방식보다 더 뛰어난 분석 방식임을 밝혔다.

상기의 연구결과는 향후 다양한 전문 학술분야에 대한 후속 연구에서 중요한 키워드 분석 방식으로 널리 활용될 것으로 사료된다. 하지만 Egbert & Biber(2019)가 사용빈도를 완전히 배제한 분포도 기반의 키워드 분석 방식에 잠재적인 문제점이 있음을 스스로 인정하였을 뿐만 아니라 Gries(2021) 역시 사용빈도를 완전히 배제한 분포도 기반의 키워드 분석이라는 가정에서 사용빈도의 완전한 배제는 불가능하다는 점을 지적하였다. 더 나아가 Gries(2021)는 궁극적으로 최적의 키워드는 어휘의 사용빈도뿐만 아니라 분포도를 모두 고려하여 추출되어야 한다고 주장하였다. 본 연구에서는 Gries(2021)에서 새롭게 제시한 이차원적 키워드 분석 방식의 효용성을 차후 연구과제로 돌리고자 한다.

REFERENCES

[1] K. S. Folse. (2004). *Vocabulary Myths: Applying Second Language Research to Classroom Teaching*. Ann Arbor: University of Michigan Press.

[2] M. Lewis. (1993). *The Lexical Approach*. Hove: Language Teaching Publications.

[3] M. Lewis. (2000). *Teaching Collocation: Further Developments in the Lexical Approach*. Hove: Language Teaching Publications.

[4] I. S. P. Nation. (1994). *New Ways in Teaching Vocabulary*. Alexandria, Va: TESOL.

[5] R. Waring & P. Nation. (2004). Second Language Reading and Incidental Vocabulary Learning. *Angles on the English Speaking World, 4*, 97-110.

[6] A. Zareva, P. Schwanenflugel & Y. Nikolova. (2005). Relationship between Lexical Competence and Language Proficiency: Variable Sensitivity. *Studies in Second Language Acquisition, 27(4)*, 567-595.

[7] I. S. P. Nation. (1990). *Teaching and learning vocabulary*. Boston: Heinle and Heinle.

[8] P. Baker. (2004). Querying keywords: questions of difference, frequency and sense in keywords analysis. *Journal of English linguistics, 2(4)*, 346-359.

[9] J. Egbert & D. Biber. (2019). Incorporating text dispersion into keyword analyses. *Corpora, 14(1)*, 77-104.

[10] S. T. Gries. (2021). A new approach to (key) keywords analysis: using frequency, and now also dispersion. *Research in Corpus Linguistics, 9(2)*, 1-33.

[11] M. Bondi & M. Scott. (2010). *Keyness in texts*. Amsterdam: John Benjamins Publishing Company.

[12] L. Grabowsky. (2015). Keywords and Lexical Bundles within English Pharmaceutical Discourse: A Corpus-driven Description. *English for Specific Purposes, 38*, 28-33.

[13] S. E. Jhang & S. M. Lee. (2013). Visualization of Collocational Networks: Maritime English Keywords. *Language Research, 49(3)*, 781-802.

[14] M. Scott & C. Tribble. (2006). *Textual Patterns: Keyword and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

[15] M. Y. Yoo. (2019). A Study on the Meaning and Collocation for Make Used in NIV English Bible. *The Journal of Humanities and Social Sciences 21, 10(3)*, 633-644.

[16] M. Scott. (2020). *WordSmith Tools Version 8.0*. Liverpool: Lexical Analysis Software.

[17] T. Dunning. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19(1)*, 61-74.

[18] S. E. Jhang & S. M. Lee. (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.

[19] P. Baker. (2010). Corpus methods in linguistics in L. Litosseliti (ed.) *Research Methods in Linguistics*, PP. 95-113. New York: Continuum.

[20] J. Culpeper. (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.

[21] M. Scott. (2016). *WordSmith Tools Manual*. Liverpool: Lexical Analysis Software.

하 명 호(Myoungho Ha) [정회원]

· 1991년 2월 : 부산대학교 영어영문학과 (문학사)
· 1995년 2월 : 부산대학교 영어영문학과 (문학석사)
· 2001년 8월 : 부산대학교 영어영문학과 (문학박사)
· 2015년 3월 ~ 현재 : 신라대학교 교양과정대학 교수
· 관심분야 : 코퍼스, 교육
· E-Mail : hadash21@silla.ac.kr