http://dx.doi.org/10.5392/JKCA.2013.13.09.440

코퍼스에 기반한 문학텍스트 분석

Corpus-Based Literary Analysis

하명정
상명대학교 영어영문학과

Myung-Jeong Ha(mjha@smu.ac.kr)

요약

코퍼스 언어학이 연구방법의 한 분야로서 최근 그 입지를 급격하게 넓혀온 가운데, 언어학적 현상과 함께 문학텍스트의 이해를 깊게 하는데 기여를 해 왔다. 최근 코퍼스 언어학의 급속한 저변확대에도 불구하고 문학텍스트 코퍼스를 기반으로 한 고전 및 문학작품의 재해석에 대한 시도는 국내언어학계에서 매우 미미한 실정에 머물러 있다. 이에 본 연구는 코퍼스 언어학의 분석도구인 컴퓨터 콘코던스 프로그램인 워드스미스를 이용하여 방대한 전자텍스트로 이루어져 있는 문학작품의 문체적 특성과 주요테마를 조사하고자 하였다. 특히 본 연구는 텍스트의 주요한 특성을 나타내는 키워드(keyword)에 초점을 두고 세익스피어의 비극작품인 로미오와 줄리엣을 코퍼스 언어학적 분석기법으로 접근하여 작품세계를 재조명하여 학문적 의의가 크다고 생각되며 앞으로 관련된 후속연구가 이어질 것으로 기대된다.

■ 중심어 : | 코퍼스 기반 분석 | 키워드 분석 | 코퍼스 문체론 | 문학텍스트 |

Abstract

Recently corpus linguistic analyses enable researchers to examine meanings and structural features of data, that is not detected intuitively. While the potential of corpus linguistic techniques has been established and demonstrated for non-literary data, corpus stylistic analyses have been rarely performed in terms of the analysis of literature. Specifically this paper explores keywords and their role in text analysis, which is primary part of corpus linguistic analyses. This paper focuses on the application of techniques from corpus linguistics and the interpretation of results. This paper addresses the question of what is to be gained from keyword analysis by scrutinizing keywords in Shakespeare's Romeo and Juliet.

■ keyword : | Corpus-based Analysis | Keyword Analysis | Corpus Stylistics | Literary Text |

I. 연구배경 및 연구목적

코퍼스에 기반한 언어학적 분석은 직관적인 판단이 어려운 텍스트의 의미와 구조적인 특징을 밝히는데 유용하다. 이러한 코퍼스 기반 분석이 논픽션 텍스트의 분석에는 널리 사용되어 왔으나, 픽션 텍스트의 분석에는 거의 적용되지 않았다. 이에 대한 예외적인 연구로는 Tabata [1]와 Stubbs[2] 등을 꼽을 수 있다. 전통적인 문체론적 분석은 코퍼스 언어학적 분석기법을 전혀 사용하지 않고 문학텍스트를 분석한다(e.g., [3]).

* 본 연구는 2013학년도 상명대학교 교내연구비를 지원받아 수행되었음.
* 본 논문은 한국콘텐츠학회 JCCC 2013 제1회 융합콘텐츠 제주학술대회 우수논문입니다.
접수일자 : 2013년 08월 06일
심사완료일 : 2013년 08월 30일
수정일자 : 2013년 08월 26일
교신저자 : 하명정, e-mail : mjha@smu.ac.kr

본 연구에서 코퍼스 문체론(corpus stylistics)이란 코퍼스 언어학(corpus linguistics)과 Halliday[4]가 텍스트 분석에서 전제로 한 문체론(stylistics)의 조합을 의미한다. 전통적인 접근법과 비교하여 코퍼스 언어학적 분석 기법은 다음과 같은 두 가지 장점이 있다:

1) 직관적으로 파악하기 불가능했던 난해한 텍스트의 의미 파악; 2) 객관적으로 존재하는 언어학적 패턴을 근거로 한, 정의상 주관적인 텍스트의 분석

이러한 두 가지 장점은 코퍼스 문체론이 연구자의 직관으로는 보이지 않는 언어학적 패턴의 식별을 가능하게 하는 컴퓨터에 기반한 분석이기 때문이다. 본 연구에서 코퍼스 문체론의 이론적인 근거는 Sinclair[5]에 있다. 코퍼스 문체론의 가장 중요한 원리는 언어학적 자료에서 빈도수(frequency)가 중요성(significance)과 같은 의미로 사용된다는 것이다. 이는 코퍼스 언어학이 텍스트에서 특별히 빈번하게 나타나는 언어적 특질을 그 텍스트의 담화구조나 의미에 중요하다고 가정한다는 것을 뜻한다.

이에 본 연구는 최근 문학텍스트의 코퍼스 언어학적 분석기법 경향을 반영하여 코퍼스 문체론의 주요한 부분인 키워드 분석(keyword analysis)을 이용하여 셰익스피어의 작품인 로미오와 줄리엣에 나오는 주요 인물들의 언어를 분석하고자 한다.

II. 이론적 배경

1. 코퍼스 문체론(Corpus stylistics)

텍스트의 부분 발췌문을 수작업으로 직관에 의존하던 종래의 언어학적 패턴 분석에서 벗어나 코퍼스 문체론은 방대한 양의 텍스트를 어휘적, 구문적, 문법적인 패턴 분석이 가능하게 한다. Jakobson[6]에 따르면, 텍스트에서 빈번하게 나타나는 문구 분석과 같은 텍스트의 종적 통합축(syntagmatic axis)이 분석의 핵심이다.

코퍼스 분석기법을 이용하여 많은 논픽션 텍스트의 분석이 행해져 왔으나 문학 텍스트의 분석은 흔치 않다. 예외적으로 Burrows[7]가 제인 오스틴(Jane Austin)의 등장인물들이 독특하게 사용하는 말(idiolect)을 분석했고, Tabata[1]가 찰스 디킨스(Charles Dickens)의 소설로 구축된 코퍼스를 분석하였다.

코퍼스 문체론에서 지대한 관심 중 하나는 작가들의 특유한 언어를 분석하는 것이었다. 예를 들면, Tabata[8]는 디킨스 소설 10편에서 발췌한 코퍼스를 분석하였다. 디킨스의 소설은 20년 이내에 씌어지고 각각 약 2만단어로 이루어져 있다. Tabata는 디킨스 소설에서 가장 빈번하게 사용된 어휘에 관심이 있었고 이들 어휘는 대부분이 문법적 어휘들로 디킨스의 문체가 변화해 간 발달사를 보여 주었다. Tabata의 연구는 디킨스 언어에 대해 새롭고 흥미로운 시각을 제공하고 또한 이러한 새로운 시각이 다른 연구에서 텍스트 각각의 의미에 대한 통찰력을 얻는데에 사용된다.

Tabata[1]와 마찬가지로 Mahlberg[9]는 디킨스의 소설 23편으로 구축한 코퍼스를 분석하였는데, 키클러스터(key cluster)로 불리는 서너 개의 단어로 이루어진 구를 추출해 낸다. 키클러스터는 참조 코퍼스보다 분석 대상인 코퍼스에서 현저하게 많이 나타나는데 코퍼스 분석 프로그램의 하나인 워드스미스(WordSmith Tools)에 의해 추출된 구이다.

2. 키워드(Keywords)의 정의

코퍼스 언어학에서 키워드(keyword)는 사회적, 문화적, 정치적인 중대성을 담기 때문에 핵심(key)으로 간주되는 어휘 항목과는 엄연히 구분되어야 한다. 본 연구에서 키워드의 개념은 통계적으로 유의미한 어휘 항목을 의미한다. 텍스트에 대한 통계학적인 특징을 지우는 어휘가 키워드인데 새로운 개념이 아니라 50년의 역사를 가지고 있다[10].

키워드는 단순히 통계적으로 유의미함을 떠나 텍스트의 문체와 밀접한 상관관계에 있다. 요컨대, 키워드란 빈도수가 통계적으로 유의미하게 차별화되는 “스타일 마커(style marker)”라고 할 수 있다[11]. 코퍼스 언어학에서 문체는 빈도수(frequency), 확률(probability)과 규범(norm)이라는 용어들로 정의가능하다. 요컨대 문체는 특정한 문맥에서 언어 항목의 빈도수와 관련이 있고 이를 문맥적 확률(contextual probabilities)이라고 칭한다[11].

아래의 [그림 1]이 코퍼스 언어학에서 사용되는 키워드의 특징을 제시하고 있다.

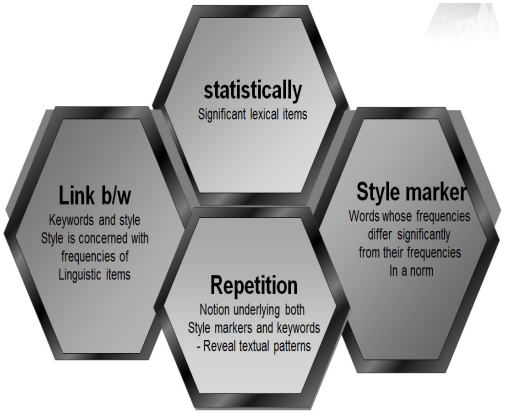

그림 1. 키워드의 특징

[그림 1]에서 보여 주는 바와 같이 키워드란 기준(norm)이 되는 텍스트에 나타나는 단어의 빈도수와 관심대상이 되는 텍스트에 나타나는 특정 단어의 빈도수가 현저하게 큰 차이가 날 경우, 이들 단어를 일컫는다. 따라서 키워드는 텍스트에서 반복적으로 나타나고 궁극적으로 텍스트의 패턴을 드러낸다.

키니스(keyness)의 분석은 비교되는 대상인 기준(norm)에 비해 상대적으로 얼마나 통계적으로 유의미하게 나타나는가의 문제이다[10]. 관련되는 통계적인 기법은 카이 스퀘어(chi-square)나 로그 라이클리후드(log likelihood)테스트가 있다. 키워드에 기반한 언어학적 분석은 직관적인 판단이 어려운 텍스트의 의미와 구조적인 특징을 밝히는데 매우 유용하다.

키워드 분석은 Mike Scott[12]가 개발한 워드스미스(WordSmith Tools)의 일부기능인 KeyWords를 통해 더욱더 보편화되었다. 워드스미스로 키워드를 식별하기 위해서, 분석 대상인 코퍼스의 워드 리스트와 참조 코퍼스(reference corpus)의 워드 리스트간의 통계적 비교 분석이 용이해졌다.

3. 문학에서의 키워드

어휘의 패턴(lexical patterns)은 '특정언어에서 구성의 형태(forms of organization in the language)'[13]로 텍스트의 내용과 구조를 동시에 결정한다. 따라서 텍스트의 어휘 패턴을 파악하는 것이 텍스트가 지니는 의미를 알아내는데 도움을 준다. 이러한 패턴의 식별에 종종 컴퓨터로 자동화된 분석이 사용되는데 이를 정량적 키워드(quantitative keywords)라고 한다[14].

언어학적 연구에서 긴 역사를 지는 키워드 식별은 애초에 직관적으로 이루어졌으나 요즘은 자동화된 과정을 거친다. 기저에 깔린 가정은 키워드가 언어나 사회의 중요한 개념을 식별하고 설명하는데 유용하다는 것이다. 키워드의 정량적인 식별법은 데이터 셋에서 키워드를 자동적으로 추출하는 통계학적 계산을 이용한 소프트웨어에 기반을 둔다. 여기에 적합한 키워드의 정의는 기준이 되는 참조 코퍼스(reference corpus)와 비교해서 특정 코퍼스에서 통계적으로 현저하게 많이 나타나는 단어들이다[15].

정량적 키워드의 장점은 다음과 같다: 첫 째, 정량적인 키워드는 분석 중인 텍스트의 중요한 개념에 주목하게 한다[16]. 이는 키워드 분석이 서로 다른 데이터 셋의 어휘적 차이에 주목하기 때문이다. 둘 째, 키워드가 데이터에 나타난 주제를 가리키고 주제사이의 관계를 드러내는 점에 의거하면 사실상 텍스트의 연어(collocation)라고 할 수 있다[17]. Mason과 Platt[17]은 T*extual Collocate*를 정의하는데 일반적인 코퍼스(general corpus)의 어휘 빈도수에 비교하여 분석하는 텍스트의 어휘와 그 어휘에 인접한 단어들이 결합된 어휘 공간(lexical span)의 빈도수가 현저하게 높은 경우 이를 textual collocate로 정의하였다. Textual collocate는 Scott[15]의 키워드와 개념적으로 매우 유사하다. Scott와 Tribble[18]은 키워드로 텍스트의 내용을 알려주는 어휘 패턴을 밝힐 수 있다고 주장하고, 이러한 워드스미스 프로그램을 이용한 어휘 패턴의 발견이 보다 심화된 텍스트 분석으로 이어지는 초석이 된다고 주장하였다.

문체론에서 키워드는 텍스트의 문학적 의미를 드러내는 지표기능을 한다고 볼 수 있다. 예를 들면, Culpeper[19]는 키워드 분석을 하여 세익스피어의 작품 로미오와 줄리엣의 주연인물들의 특징을 깊이있게 조사하였다. Culpeper는 이 연구에서 주연인물들이 일인

칭 단수와 복수 대명사를 상황에 맞게 사용하여 자신의 사회적 지위와 성격을 드러냈다고 보고하였다. Culpeper는 후속연구로서 로미오와 줄리엣의 키워드 분석뿐 아니라 키워드의 의미론적 범주와 문법을 조사하였다.

4. 키워드분석(Keyword analysis)의 고려할 점

키워드분석을 함에 있어서 고려해야 할 점들이 다음과 같이 몇 가지가 있다.

먼저, 분석하고자 하는 텍스트와의 비교를 위한 데이터(즉 reference corpus)의 선택이 중요하다. 데이터는 일반적으로 수천 개의 어휘로 구성된 사이즈가 큰 데이터가 선호된다[18]. Sardinha와 Barbara[20]는 참조 코퍼스의 사이즈가 크기와 키워드의 개수가 비례한다고 제안하고 참조 코퍼스가 분석대상인 텍스트의 적어도 5배 이상의 크기가 되어야 한다고 주장한다.

분석할 대상인 텍스트와 장르면에서 밀접한 관계가 있는 참조 코퍼스일수록 키워드분석이 흥미로운 결과를 보여줄 가능성이 크다. 즉 참조 코퍼스의 선택이 키워드분석 결과에 큰 영향을 미친다고 할 수 있다. 단순히 BNC(British National Corpus)와 같이 여러 가지 장르로 구성된 방대한 분량의 텍스트를 참조코퍼스로 활용한다면 BNC의 특정 텍스트는 분석하고자 하는 텍스트와 밀접한 관계가 있고, 다른 장르로 된 텍스트는 상대적으로 먼 관계를 가지기 때문에 이러한 관계의 다양성이 추출되는 키워드에 영향을 미칠 것이다.

따라서 본 연구는 로미오와 줄리엣이 세익스피어 작품 중 주요한 비극작품임에 착안하여 세익스피어의 다른 비극작품 5편(햄릿, 줄리어스 시저, 리어왕, 오델로, 멕베드)을 참조코퍼스로 구축하여 장르적인 공통점을 추구하였다. 본 연구를 Scott와 Tribble[18]이 시행한 분석과 비교했을 때, 그들은 세익스피어의 모든 비극과 희극을 함께 묶어 참조코퍼스로 사용해서 다소 다른 키워드를 산출하였다.

본 연구의 분석도구로 사용하는 워드스미스의 키워드 프로그램을 사용하기 위해서는 키워드의 최소 빈도 구분점을 정해야 한다. 일반적으로 10으로 최소 빈도 구분점을 정하지만 본 연구에서 키워드에 대한 최소 빈도수는 18로 정하여 상대적으로 많은 추출된 키워드 수를 조절하였다. 다음으로 고려할 척도는 키워드의 유용성의 정도를 계산하는 통계적 유의성 테스트이다. Rason[21]을 따라 본 연구에서는 통계적 유의수준(p-value)을 0.05로 지정하고 카이 스퀘어 검정 대신 로그 라이클리후드(log-likelihood) 테스트를 사용하였다.

Ⅲ. 데이터 분석과정

본 연구에서는 분구텐베르크 프로젝트(http://www.gutenberg.org/ebooks)에서 세익스피어 작품인 로미오와 줄리엣 텍스트를 다운로드받아서 분석대상 코퍼스를 구축하였다. 또한 세익스피어의 대표적인 비극작품 5편, 즉 햄릿, 줄리어스 시저, 리어왕, 오델로, 멕베드를 각각 다운로드 받은 후에 로미오와 줄리엣과 비교할 참조코퍼스를 구축하였다. 모든 파일은 유니코드로 변환시켜 워드스미스에서 분석가능하게 만든 후, 워드 스미스에 있는 키워드 프로그램 도구를 사용하여 키워드분석을 시행하였다. 키워드 분석을 위하여 먼저 로미오와 줄리엣 코퍼스와 참조 코퍼스의 워드리스트를 각각 산출하였다.

아래의 [그림 2]는 로미오와 줄리엣의 워드리스트를 산출한 결과를 보여 준다.

N	Word	Freq.	%	Texts	%
1	AND	719	2.77	1	100.00
2	THE	679	2.62	1	100.00
3	I	586	2.26	1	100.00
4	TO	574	2.21	1	100.00
5	A	470	1.81	1	100.00
6	OF	400	1.54	1	100.00
7	MY	360	1.39	1	100.00
8	THAT	347	1.34	1	100.00
9	IS	344	1.33	1	100.00
10	IN	320	1.23	1	100.00
11	YOU	291	1.12	1	100.00
12	THOU	277	1.07	1	100.00
13	ME	265	1.02	1	100.00
14	NOT	260	1.00	1	100.00
15	WITH	255	0.98	1	100.00
16	IT	228	0.88	1	100.00
17	THIS	226	0.87	1	100.00
18	FOR	225	0.87	1	100.00
19	BE	215	0.83	1	100.00
20	BUT	183	0.71	1	100.00
21	WHAT	165	0.64	1	100.00
22	THY	164	0.63	1	100.00

그림 2. 로미오와 줄리엣의 워드리스트

각 코퍼스의 워드리스트를 산출한 후에 각각의 워드리스트를 비교하는 키워드 분석결과 아래와 같은 결과를 얻었다. [그림 3]은 로미오와 줄리엣의 키워드리스트를 보여 준다.

N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness
1	A	470	1.81	1,896	1.50	13.25
2	IS	344	1.33	1,293	1.02	17.70
3	THOU	277	1.07	664	0.53	88.95
4	ME	265	1.02	1,044	0.83	9.30
5	WITH	255	0.98	1,010	0.80	8.52
6	FOR	225	0.87	942	0.75	4.13
7	THY	164	0.63	441	0.35	38.54
8	ROM	163	0.63	0		578.19
9	HER	156	0.60	504	0.40	18.81
10	O	154	0.59	351	0.28	55.47
11	NURSE	150	0.58	1		520.35
12	LOVE	140	0.54	153	0.12	147.72
13	THEE	138	0.53	389	0.31	27.97
14	ROMEO	136	0.52	0		482.29
15	JUL	117	0.45	0		414.84
16	SHE	114	0.44	274	0.22	36.21
17	FRIAR	92	0.35	0		326.13
18	AN	86	0.33	216	0.17	24.41
19	NIGHT	83	0.32	202	0.16	25.54
20	HERE	80	0.31	213	0.17	19.31
21	GO	77	0.30	243	0.19	10.26
22	I'LL	72	0.28	140	0.11	35.62
23	MAN	72	0.28	263	0.21	4.47
24	DEATH	71	0.27	125	0.10	41.53

그림 3. 로미오와 줄리엣의 키워드 리스트

키워드 리스트를 산출한 후에 각 키워드의 상황문맥을 콘코던스 라인을 추출하여 하나씩 세밀히 조사하였다.

Ⅳ. 분석결과 및 논의

1. 키워드의 종류

워드스미드에서 추출한 키워드 중에 관사와 같은 기능어를 제외하고 극중 등장인물의 대사를 가리키는 인물 이름은 분석에서 제외하였다. [표 1]은 최소 빈도수가 18이상인 키워드를 정리한 결과를 제시한다.

표 1. 로미오와 줄리엣의 키워드리스트

Is (344)	Thou (277)	Me (265)	With (255)	For (225)	Thy (164)
Her (156)	O (154)	Nurse (150)	Love (140)	Thee (138)	Romeo (136)
She (114)	Friar (92)	Night (83)	Here (80)	Go (77)	I' ll (72)
Man (72)	Death (71)	Lady (68)	Juliet (58)	Up (57)	Too (57)
Tybalt (55)	Art (55)	Cap (53)	Dead (49)	Day (48)	Wife (47)
Doth (47)	Give (47)	Tell (45)	Fair (44)	Prince (36)	Paris (35)
Sweet (35)	Capulet (33)	Gone (33)	Old (33)	God (32)	Montague (31)
Ay (31)	Look (31)	Wilt (30)	Mother (30)	Light (29)	Stay (28)
Heaven (28)	House (28)	Hast (28)	Dear (27)	Bed (25)	Young (24)
Find (24)	Name (24)	Die (23)	Marry (23)	Face (23)	Madam (23)
Watch (22)	Bid (22)	Holy (21)	Lie (21)	Tears (20)	Ah (18)

* ()안은 발생빈도수임.

첫 째, 여러 개의 고유명사가 키워드리스트에 포함된 것을 알 수 있다. [표 1]에서 빈도수를 살펴보면 Romeo는 136번, Friar는 92번, 그리고 Juliet는 58번 언급되었다. 이들 고유명사는 극중 다른 인물들에 의해 언급된 것이다. 따라서 Romeo가 최다 빈도수(136번)를 기록하고 있는 점에서 극중에서 가장 중요한 역할을 차지하고 있음을 알 수 있다.

둘 째, 작품의 중요한 테마(theme)를 반영하는 키워드를 발견할 수 있다. [표 1]에서 love(140), night (83), death (71), light (29) 등이 로미오와 줄리엣을 특징짓는 주요한 키워드들이다. 이러한 키워드는 Phillips[13]가 사용한 용어, 즉 텍스트의 내용(aboutness)과 밀접하게 연관된다.

반면 흥미롭게도 텍스트의 내용과 직접 연관이 없는데도 불구하고 극의 중요성을 반영할 만한 키워드 몇 가지를 볼 수 있다. 표 1에 제시된 감탄사 O(154), Ah(18)와 대명사 thou(277), 동사 art(55)와 wilt(30)가 이에 해당한다. 이들 키워드는 빈도수가 현저히 높게 나타나기 때문에 텍스트의 내용 즉 aboutness와 직접 관련은 없는 반면, 텍스트의 형식과 관련된 문체(style)적 특징을 드러낸다고 볼 수 있다. 요컨대, 텍스트의 내용과 관련된 키워드는 개방부류 단어(open-class word)인 반면 문체를 특징짓는 키워드는 폐쇄부류 단어(closed-class word)로 문법 단어(grammatical word)인 경향이 있다[10]. 예를 들어, Art는 로미오와 줄리엣에서 명사로 쓰인 적이 없고 대명사 thou(you의 고어)뒤에 오는 2인칭 단수 be의 고어형으로 쓰였다. 이는 극이 표방하는 테마의 친밀성(intimate nature)을 드러낸다고 볼 수 있다.

더욱 흥미로운 것은 O나 Ah와 같은 감탄사의 빈도수가 상당히 높은 점이다. 감탄사가 키워드로 추출된 원인을 조사하기 위하여 문맥을 파악할 수 있는 키워드의 콘코던스 라인을 살펴 본 결과 다음과 같은 점이 발견되었다. [표 2]에서처럼 로미오와 줄리엣에 사용된 감탄사 Ah는 감탄사 O보다 다소 부정적인 상황문맥에서 사용되었음이 드러났다.

표 2. 감탄사 Ah와 O가 쓰인 콘코던스라인 비교

감탄사	Ah	O
문맥1	Ah, word ill urg'd to one that is so ill!	O holy friar, O, tell me, holy friar Where is my lady's lord, where's Romeo?
문맥2	Ah, weraday! he's dead, he's dead, he's dead!	O Lord, I could have stay'd here all the night To hear good counsel. O, what learning is! My lord, I'll tell my lady you will come.
문맥3	Ah, where's my man? Give me some aqua vitae. These griefs, these woes, these sorrows make me old.	O, now be gone! More light and light it grows.
문맥4	Ah, poor my lord, what tongue shall smooth thy name When I, thy three-hours wife, have mangled it?	O, think'st thou we shall ever meet again?
문맥5	Ah sir! ah sir! Well, death's the end of all.	O, he's a lovely gentleman!

극에서 감탄사 O는 154회 나타나는데 로미오와 줄리엣 코퍼스에서 0.59%를 차지하고 참조코퍼스에는 351번 나타나 참조코퍼스의 0.28%를 차지하였다. 즉 로미오와 줄리엣에서 감탄사 O가 사용된 비율이 참조코퍼스에 비해 2배로 높기 때문에 키워드로 인식된 것이다.

2. 키워드의 분포

키워드는 텍스트에 존재하므로 텍스트에서 어디에 위치하고 키워드끼리 어떤 연관성이 있는지가 하나의 이슈가 된다. 다음의 [그림 4]는 키니스(keyness)의 순서로 추출된 로미오와 줄리엣의 키워드를 보여주고 있다.

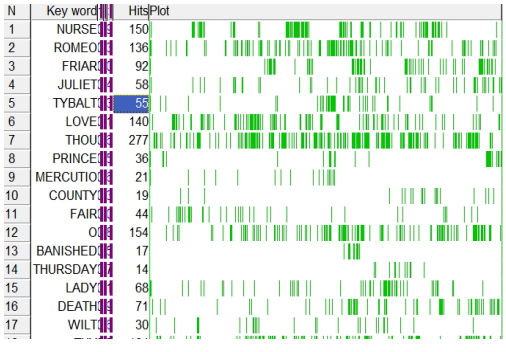

그림 4. 로미오와 줄리엣의 키워드 플롯

[그림 4]에서 오른쪽에 위치한 점선 마크가 극 전체에 산재해 있는 각 키워드의 분포를 나타내고 있다. 예를 들면, 키워드 love는 극의 후반부보다 전반부에 걸쳐서 퍼져 있는 반면, 그림의 하단에 위치한 death는 발생 빈도수가 후반부로 갈수록 많아짐을 알 수 있다. 그리고 감탄사 O는 비극으로 치닫는 후반부로 가까워질수록 증가추세에 있고, banished와 Thursday는 극의 중반부에 갑자기 나타나고 있다.

키워드는 전체 텍스트에서 분포하는 상태에 따라 글로벌 키워드(global keyword)와 로컬 키워드(local keyword)의 두 가지 종류로 구분할 수 있다[18]. [그림 4]의 키워드 분포에서 대명사 thou는 텍스트 전반에 걸쳐 비교적 균등하게 분포되어 있으므로 글로벌 키워드로 볼 수 있고, 반면에 banished는 로컬 키워드로 간주된다. 로미오와 줄리엣에서 실제로 추방(banishment)은 3막 2장에서 이슈화된다. 요컨대 키워드는 텍스트에서 전체적으로 골고루 분포하는 양상을 띠거나 국부적으로 한 곳에 집중된 양상을 보이는 경향이 있다.

V. 결론

본 연구는 코퍼스 문체론에서 중추적인 부분을 차지하는 키워드 분석을 통하여 세익스피어의 대표적인 비극작품인 로미오와 줄리엣의 문체적 특징을 조사하고자 하였다. 키워드 분석의 기본원리는 두 개의 코퍼스를 비교하여 분석대상인 텍스트에서 빈도수가 현저하게 높은 어휘를 식별하는 것이다. 기존의 문학텍스트 분석과는 달리 통계학에 기반을 둔 정량적 분석을 시도하여 코퍼스 언어학적 분석이 문체론의 연구에 어떻게 활용될 수 있는지를 보여 주고자 하였다. 또한 최근 전자 텍스트와 콘텐츠로 구성된 전자책이 인쇄책의 개념을 급속도로 계승, 변형, 확장해 나가는 시점[22]에서 문학텍스트의 코퍼스적 접근이 가지는 함의를 간과할 수 없다고 본다.

따라서 본 연구의 의의는 문학작품의 분석과 나아가 교수법의 제시에 있어서 전자화된 텍스트 뭉치인 문학코퍼스를 활용하는 실례를 자동화된 데이터 추출과정

으로 실증적으로 보여줌으로써 문학교육의 새로운 지평을 여는데 일조하였음에 있다고 본다. 교수법과 관련해서 본 연구에서 소개된 콘코던스 프로그램인 워드스미드는 교수자들이 텍스트의 지배적인 패턴을 발견하는데 유용할 뿐만 아니라 지배적인 패턴내에 존재하는 작은 언어학적 변이들을 관찰하는 데에도 도움이 될 것이다. 또한 문학 지식의 발달이 학습자의 문학적 자각(literary awareness)과 무관하지 않음[23]을 상기할 때, 문학 코퍼스에서 관찰되는 언어패턴의 인식과 주목(noticing)이 새로운 문학교육의 고안에 많은 시사점을 준다고 본다.

앞서 기술하였듯이, 코퍼스 언어학에서 키워드의 개념이 문체의 개념과 밀접한 관련이 있음을 상정하고 문학작품의 분석을 통하여 키워드 분석의 유용성을 밝히고 선행연구에서 언급되지 않은 면을 보완하고자 하였다. 이에 입각하여 키워드 분석을 시행하기 위한 고려사항들 즉 참조 코퍼스의 특성과 최소 빈도수의 한계점 설정에 대하여 먼저 리뷰하였다. 그리고 두 가지 종류의 키워드를 분별하여 내용지표(aboutness indicator 즉 love, death, night)와 형식지표(stylistic indicator 즉 thou, art, wilt)를 추출하였고, 마지막으로 로미오와 줄리엣 코퍼스에서 키워드의 분포상태를 조사함으로써 키워드의 분포가 작품의 테마나 극중전개와 어떤 연관이 있는지를 살펴 보았다.

따라서 선행연구[10][19]와 더불어 본 연구는 전통적으로 행해졌던 질적 분석과 비교해서 다음과 같은 키워드 분석의 장점을 부각시키고자 하였다.

첫 째, 키워드 분석은 연구자의 직관만으로는 간과하기 쉬운 명확하게 드러나지 않는 언어학적 자질 및 특성을 밝히는데 기여할 수 있다. 둘 째, 텍스트의 어느 부분에 주목할지 혹은 어떤 언어적 자질에 관심을 기울일 지에 대하여 직관에 의존하지 않고 키워드 분석으로 어휘의 패턴을 밝힐 수 있다. 이는 키워드 분석이 기본적으로 반복성에 기반을 둔 통계학에 기초하기 때문이다.

문학텍스트의 의미를 재해석함에 있어서 키워드 분석이 더욱 공고해지기 위해서는 앞으로 후속연구에서는 키워드 분석과 더불어 문법분석(part-of-speech analysis)과 의미론적 분석(semantic analysis)이 병행되어야 할 것으로 생각된다[24]. 특히 텍스트의 의미론적 분석은 본 연구에서 단어에 국한된 키워드 분석의 제한점을 보완하여 줄 것이다. 또한 키워드 분석이 텍스트간의 유사점은 배제하고 순수하게 차이점에만 주목하는 본연의 한계성을 지닌 것도 차후 연구를 수행함에 있어 유의해야 할 점으로 남는다.

참 고 문 헌

[1] T. Tabata, Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution. In T. Saito, J. Nakamura, and S. Yamazaki (eds.). English corpus linguistics in Japan. Rodopi, Amsterdam and New York, pp.165-182, 2002.

[2] M. Stubbs, "Conrad in the computer: Examples of quantitative stylistic methods," Language and Literature, Vol.1, No.5, pp.5-24, 2005.

[3] M. H. Short, "Discourse Analysis and the Analysis of Drama," Applied Linguistics, Vol.2, No.2, pp.180-202, 1981.

[4] M. A. K. Halliday, *Linguistic function and literary style: An inquiry into the language of William Golding's The Inheritors,* In Literary style: A symposium. London & New York, Oxford University Press, pp.330-365. 1971.

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

[6] R. Jakobson, Closing Statement: Linguistics and Poetics. In T.A. SEBEOK (ed.), *Style in Language*. Cambridge, MA: MIT 31971, pp.350-377, 1958.

[7] J. F. Burrows, *Computation into criticism: A study of Jane Austen's novels and an experiment in method.* Oxford: Clarendon Press, 1987.

[8] T. Tabata, "Dickens Narrative Style: A

Statistical Approach to Chronological Variation," RISSH, Vol.30, pp.165-182, 1994.

[9] M. Mahlberg, "Clusters, key clusters and local textual functions in Dickens," Corpora, Vol.2, No.1, pp.1-31, 2007.

[10] J. Culpeper, "Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet," International Journal of Corpus Linguistics, Vol.14, No.1, pp.29-59, 2009

[11] N. E. Enkvist, M., Gregory, and J. Spencer, *Linguistics and Style: on Defining Style: An Essay in Applied Liguistics*. Oxford University Press, 1964.

[12] M. Scott, WordSmith Tools version 5, Liverpool: Lexical Analysis Software, 2008.

[13] M. Phillips, "Lexical structure of text. No.12", English language research, 1989.

[14] B. Fischer-Starcke, *Corpus linguistics in literary analysis: Jane Austen and her contemporaries*. Continuum, 2010.

[15] M. Scott, "PC analysis of key words - and key key words," System, Vol.25, No.2, pp.233-245, 1997.

[16] P. Baker, "Querying Keywords Questions of Difference, Frequency, and Sense in Keywords Analysis," Journal of English Linguistics, Vol.32, No.4, pp.346-359, 2004.

[17] M. A. Oliver and R. Platt, "Embracing a new creed: lexical patterning and the encoding of ideology," College Literature, Vol.33, No.2, pp.154-170, 2006.

[18] M. Scott and C. Tribble, *Textual patterns: Key words and corpus analysis in language education*. John Benjamins Publishing, 2006.

[19] J. Culpeper, Computers, language and characterisation: an analysis of six characters in Romeo and Juliet, pp.11-30, 2002. http://eprints.lancs.ac.uk/id/eprint/31705.

[20] T. B. Sardinha, and L. Barbara, Corpus linguistics. *The Handbook of Business Discourse*. p.105, 2009.

[21] P. Rayson, *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Diss. 2003.

[22] 한혜원, 박경은, "전자책 콘텐츠의 체험성과 독서경험", 한국콘텐츠학회논문지, 제11권, 제12호, pp.171-181, 2011.

[23] D. Hanauer, "Attention and Literary Education," Language Awareness, Vol.8, pp.15-26, 1999.

[24] P. Rayson, "From key words to key semantic domains," International Journal of Corpus Linguistics, Vol.13, No.4, pp.519-549, 2008.

저 자 소 개

하 명 정(Myung-Jeong Ha) 정회원

- 2005년 2월 : 서울대학교 영어교육학과(교육학석사)
- 2009년 8월 : 텍사스 오스틴 대학교 외국어교육학과(문학박사)
- 2010년 8월 ~ 2011년 2월: 인하대학교 영어교육과 강의전담계약강사
- 2011년 3월 ~ 2012년 2월 : 청주대학교 영어영문학과 교수
- 2012년 3월 ~ 현재 : 상명대학교 영어영문학과 교수

<관심분야> : 멀티미디어 외국어교육, 코퍼스 언어학, 컴퓨터 보조 외국어교육