문학 코퍼스의 구성과 활용 방안

김병선 (한국학중앙연구원)

1. 어문학 연구와 언어적 코퍼스

1998년부터 시작된『21세기 세종계획』프로젝트에 의해 한국의 학계에서도 언어 코퍼스(linguistic corpus)에 대한 관심이 매우 높아졌으며, 그 결과로 여러 종류의 큰 규모 말뭉치들이 구축되었다.[1] 말뭉치뿐만 아니라 말뭉치를 활용할 수 있는 프로그램들도 개발되었으며, 특별히 말뭉치 언어학(corpus linguistics)의 범주에 속하지 않더라도 대부분의 언어학 연구에서는 이러한 말뭉치에 대한 조사를 기본으로 하고 있다. 문학작품의 텍스트 역시 언어적 저작물인 상상적 텍스트로서 언어 코퍼스의 무시할 수 없는 부분을 차지하고 있다. 국립국어원에서는 일찍이 문학작품의 코퍼스를 축적하여『표준국어대사전』의 편찬에 활용하였고, 그러한 문학 텍스트로부터 표제어를 추출하거나 의미의 맥락을 확인하거나 사전의 의미를 기술할 때에는 용례의 제시에 활용하기도 하였다.

문학 연구자들도 문학 코퍼스에 관심을 가져서 한 작가의 어휘사전이나 용례사전의

[1] 필자는 1997년에 '21세기 세종계획'의 기획 단계에 참여하였고, 1998년부터는 구비문학 말뭉치의 구축 작업에도 참여하였다. 대규모 말뭉치로는 연세대학교의 말뭉치와 고려대학교의 말뭉치 등이 있다.

편찬에 활용한다든가, 어휘 빈도에 대한 연구의 기초자료로 활용하기도 하였지만,[2] 문학 텍스트의 코퍼스에 대한 본격적인 계량적 연구로는 아직 이어지지 못하고 있다. 기존의 코퍼스는 문학작품 중에서도 주로 소설 장르에 치우쳐 있고, 계량적 연구를 수행하기에는 표본 코퍼스의 규모가 작거나 균형성이 부족한 것들이기 때문이다. 또한 기존의 코퍼스는 문학 연구를 위한 것이기보다는 언어 연구를 위한 것으로서, 문서에 대한 마크업도 일반적인 언어 코퍼스의 차원에 머무르고 있는 실정이다.

기존의 문학 텍스트 코퍼스 중, 현대소설의 경우는 저작권 문제로 입력만 했을 뿐, 일반에 공개될 수 없는 상황이고, 저작권 문제가 해결된 신소설 같은 근대 텍스트는 자료의 신빙성에서 적지 않은 문제가 있었다. 텍스트에 대한 원본비평적인 검토가 거의 이루어지지 않은 상태였다. 따라서 필자는 문학 연구에 적용할 만한 코퍼스의 구축에 관심을 가지고 지금까지 여러 기회를 통해 문학 코퍼스를 구축해 오고 있다.

2. 문학 연구를 위한 코퍼스의 구축

2.1. 김소월 시어 색인으로부터 출발

필자는 학문의 길에 들어서면서부터 문학 연구에 컴퓨터를 활용하는 것에 대해 관심을 가지고 있었다. 1980년대 말부터 개인용 컴퓨터에 의한 문헌자료의 입력과 가공이

---

[2] 필자가 소월시 코퍼스를 바탕으로 편집한 『소월의 시어와 그 쓰임새』(한국문화사, 1994)도 문학 어휘에 대한 용례사전의 하나다.

가능하게 되면서 필자는 전공 분야인 현대시 작품의 데이터베이스 구축에 관심을 가지기 시작하였다. 일단 완전한 자료 처리의 수준은 아닐지라도 당시 사용하던 <보석글>이나 <한글> 등의 워드프로세서에서 옛한글의 입력과 출력이 어느 정도 가능했으며, 초보적인 수준은 벗어날 정도의 자료 가공도 가능한 상황이었다. 이러한 상황에서 시어의 용례색인(concordance)을 제작해 보기로 하였고, 그 첫 번째 대상으로 김소월의 시작품 전집을 선정하였다.[3]

「한글 맞춤법 통일안」이 공표되기 전인 1920년대에 발표된 김소월의 작품에는 다수의 옛한글이 포함되어 있었다. 필자는 이러한 옛한글 포함한 한국에서 사용된 문자들을 온전히 표현할 수 있는 방안을 강구하는 한편, 용례색인 및 어휘 통계 정보까지를 포함하는 결과물을 만들어낼 수 있었다. 최초에는 용례색인도 워드프로세서의 기능들을 활용하여 제작했지만, 스스로 프로그래밍 언어를 공부한 끝에 이러한 용례색인 과정을 자동으로 생성할 수 있는 프로그램(『똑똑새』)도 만들었다. 생성된 용례색인을 MS Access의 테이블로 수입하여, 각종의 정보 처리가 가능하도록 하였다.

2.2. 시작품 코퍼스에의 도전

각종의 문학 연구 자료가 영인본 형태로 활발하게 출판되던 1990년대 중반에는 『한국시사자료집성』(태학사)과 『한국시사자료대계』(한양대학교)라는 대규모의 현대시집 영인본 전집도 출간되었다. 『집성』에는 모두 46책에 228종의 시집에 포함되어 있으며, 『대

[3] 소월 시의 용례색인을 제작하게 된 데에는 수작업으로 이루어진 이상섭의 『「님의 침묵」의 어휘와 그 활용 구조』(탐구당, 1984)를 전산처리의 방법으로 재현해 보고자 하는 의도가 있었다.

계』에는 모두 39책에 189종의 시집이 포함되어 있는데, 두 전집에서 중복되는 87종의 시집을 제외하면 모두 230종의 시집이 일시에 출판된 것이다. 비록 영인본으로서 인쇄상태가 좋지는 못했지만 원본 자료를 가까이 두게 되니 필자에게는 자연 한국 현대시에 대한 코퍼스 구축의 의욕이 생겼다. 이후 4~5년에 걸쳐서 현대시집과 현대시 목록을 작성하고, 현대시 본문을 입력한 것이 본격적인 문학 코퍼스의 출발점이 된 것이다.

한국 현대시 작품은 한국학중앙연구원의 프로젝트로 데이터베이스로 구축되었다.[4] 이 데이터베이스에는 1923년부터 1950년 사이에 출판된 창작시집의 작품만을 수록하였다.『집성』과『대계」의 해당 시기의 시작품 텍스트는 다 입력하였고, 이 외에 한국학중앙연구원 도서관 소장본 김명순의『생명의 과실』(한성도서, 1925)도 추가하였다. 그 결과 모두 8,201 편의 작품에서 모두 60만 개 이상의 어휘(token)에 42,485 종의 어종(type)으로 된 현대시 코퍼스를 갖추게 되었다.

이 프로젝트의 결과물은 누리미디어(주)의 krpia를 통해서 '한국현대시어 용례사전'이라는 이름으로 서비스되고 있으며, 빈도 자료를 따로 조사하여『한국 현대시어 빈도사전』(한국문화사, 2007)이라는 종이책도 출판되었다.

필자는 이 현대시 코퍼스를 확장하여 시어의 변화상 등을 연구할 수 있도록 1950~1960년의 시 텍스트를 확보하고자 한다. 또한 각 시인별로 전집 형식의 텍스트

---

[4] 이 현대시 코퍼스에 대해서는 다음 논문 참조. 김병선,「현대시 데이터베이스 구축에 관한 연구」,『한국어문』3집, 한국정신문화연구원. 1993. pp.85-150. 김병선,「한국 현대시 데이터베이스의 구성과 그 활용방안」,『한국언어문학』제53집, 한국언어문학회. 2004. pp.513-535.

를 확보하는 것을 고려하고 있다. 후자의 경우는 시인론을 연구하는 연구자들과의 공동 작업을 통해 확보할 계획이다. 현재 필자의 현대시 코퍼스에는 포함되지 않은 시작품 텍스트도 다수 확보하고 있다. 이러한 자료는 추후 적절한 과정을 거쳐서 한국현대시 코퍼스에 수록할 계획이다.

2.3. 신소설 어휘사전의 편찬

2009년부터 2011년에는 한국학중앙연구원의 프로젝트로 신소설 어휘사전의 편찬에 도전해 보았다. 관련 연구자들과 함께 공동 연구 형태로서 20세기 초반에 출판된 신소설(창작소설을 위주로 하되 번안소설과 번역소설도 포함함.) 59편을 입력하여 어휘 용례색인을 만들었으며, 현대시 코퍼스에서 진일보하여 신소설 어휘에 대한 사전적 의미까지 정의함으로써 어휘사전 형식으로 마무리하였다.[5]

2009년(21편)		2010년(23편)		2011년(15편)	
작품명	어절수	작품명	어절수	작품명	어절수
고목화(상)	11,683	강상루	32,877	계명성	5,183
고목화(하)	9,303	강상촌	20,828	공진회	11,300
구마검	16,070	목단화	17,615	귀의성(상)	20,219

[5] 신소설 어휘사전 편찬에 대해서는 다음 논문을 참조. 김병선, 「신소설 코퍼스 구성과 어휘사전 표제어 정보의 처리」, 『신소설 어휘사전 편찬 II』, 한국학중앙연구원 어문생활사연구소 학술회의논문집, 2010. pp.5-34.

구의산(상)	12,113	설중매	12,642	귀의성(하)	17,987
구의산(하)	11,603	설중매화	12,455	금강문	29,191
두견성(상)	18,830	세검정	13,812	금국화(상)	12,348
두견성(하)	19,266	송뢰금	16,019	금국화(하)	12,923
모란병	17,702	쌍옥적	15,473	금수회의록	6,155
비파성	28,487	안의성	28,401	금의쟁성	9,707
빈상설	21,972	애국부인전	6,064	마상루	10,419
산천초목	10,566	연광정	14,131	명월정	15,107
옥호기연	8,211	요지경	17,283	비행선	17,398
우중행인	30,868	월하가인	18,143	완월루	10,662
원앙도	12,597	은세계	18,103	죽서루	7,955
자유종	6,591	재봉춘	20,645	철세계	14,433
추풍감수록	11,230	추월색	13,631		
혈의누	12,429	추천명월	12,137		
홍도화(상)	11,186	치악산(상)	21,515		
홍도화(하)	16,800	치악산(하)	18,133		
화세계	20,138	행락도	18,661		
화의혈	18,985	현미경	25,276		
		화중화	7,477		
		황금탑	12,879		
소계	326,630	소계	394,200	소계	200,998

표 1 신소설 코퍼스의 작품과 어휘수

오윤선의 신소설 목록[6] 데이터베이스 파일을 제공 받아, 이를 바탕으로 연구 대상 작품을 정리하였다. 작품 텍스트는 기존에 영인본으로 출판된『한국개화기문학총서』(아세아문화사, 1985), 및『신소설전집』(계명문화사, 1987) 등을 스캔하여 이미지 파일을 확보하였고, 여기에 포함되지 않은 판본은 국립중앙도서관의 이미지 파일을 활용하였다.

당초에 세종 코퍼스에 포함된 신소설 텍스트를 활용할 계획이었으나, 텍스트에 오류가 많은 관계로 새로 입력하였다. 이때 수작업으로 입력하지 않고, 이미지를 텍스트로 변환하는 방법을 적용하였다. 확보한 이미지 파일을 대상으로 하여, 나라지식정보(주)에 위탁하여 이미지 세그멘테이션 기법과 입력자에 의한 판독 기법을 적용하여 입력함으로써 입력 오류를 최소화하였는데, 어구나 행의 누락 없이 입력할 수 있었던 것은 이와 같은 입력 방식을 적용한 덕분이었다.

입력 자료는 용례색인 형식으로 변환하였고, 이를 다시 MS Access의 테이블로 수입하여 처리하였다. 그 결과 총 90만 어절(token)을 상회하는 코퍼스가 구축되었으며, 총 40,114 종의 어휘를 표제어로 확정하였다. 프로젝트의 결과물은 '신소설 어휘사전' 사이트에서 서비스되고 있다.[7]

---

[6] 오윤선의 「신소설 서지 데이터베이스의 분석과 그 의미」,『우리어문연구』25집, 우리어문연구회. 2005. pp.549-584.

[7] 현재는 시범 서비스 중으로 기본형 어휘에 대한 문자열 검색만 지원하고 있다. 향후에 활용형 검색 등

그림 **오류! 책갈피가 정의되어 있지 않습니다.**한국신소설어휘사전 검색화면

(newnovel.aks.ac.kr)

2.4. 현대소설 문체 연구 자료의 축적

2014년부터는 그 동안 미개척지로 남아 있는 현대소설 코퍼스의 구축에 착수하였다. 문학의 주요 장르인 소설 말뭉치가 확보된다면, 현대시와 신소설 그리고 현대소설까지를 포함한 문학 코퍼스로써 장르별 어휘의 비교 연구라든지, 산문과 운문의 문체 특성이라든지, 신소설과 현대소설의 언어적 관련성과 그 변화상 등에 대한 다양한 연구가 가능하게 된다. 따라서 100만 어절 규모의 현대소설 코퍼스 구축을 목표로 하고 이를 향후 몇 년에 걸쳐서 시기별로 작품 원문을 입력하기로 하였다.

다양한 검색 옵션을 추가할 예정이다.

우선 일차 년도에는 16명의 소설가 작품 45편을 입력하여 총 26만 어절 규모의 원시 말뭉치(raw corpus)를 구축하고 있다. 이 코퍼스는 원문 입력 및 문장 단위의 구획 및 문장 성격(서술인지 대화인지) 및 화자(작가인지 어떤 인물인지)에 대한 태그를 붙였다.

소설가	발표연도	작품명	어휘수
계용묵	1927	최서방	2749
계용묵	1928	인두지주	1637
김기진	1924	붉은 쥐	3863
김기진	1925	젊은 이상주의자의 사	3566
김동인	1919	약한 자의 슬픔	11451
김동인	1921	배따라기	3162
김동인	1922	태형	3823
김동인	1925	감자	1560
나도향	1923	행랑 자식	3169
나도향	1925	뽕	3698
나도향	1926	벙어리 삼룡이	2939
박영희	1925	사냥개	1786
박영희	1925	전투	4705
박영희	1926	지옥 순례	2095
박영희	1926	철야	1697

송영	1925	늘어 가는 무리	4123
송영	1927	군중 정류	3245
송영	1927	석공조합대표	3844
염상섭	1922	만세전	27116
유진오	1929	오월의 구직자	6133
이광수	1917	무정	80584
이태준	1929	그림자	3438
이효석	1928	도시와 유령	2999
전영택	1919	천치냐 천재냐	2637
전영택	1920	생명의 봄	12730
전영택	1920	운명	4148
전영택	1921	독약을 마시는 여인	2056
전영택	1925	화수분	2064
조명희	1925	땅 속으로	7044
조명희	1926	R군에게	4212
조명희	1926	농촌 사람들	3699
조명희	1926	저기압	1461
조명희	1927	낙동강	3121
주요섭	1921	추운 밤	2296
주요섭	1927	개밥	2649
최서해	1924	고국	1172

최서해	1925	기아와 살육	2688
최서해	1925	박돌의 죽음	2758
최서해	1925	큰물 진 뒤	3386
최서해	1927	전아사	4910
최서해	1928	갈등	7167
현진건	1921	빈처	3453
현진건	1925	B사감과 러브레터	1358
현진건	1925	불	1769
현진건	1926	고향	1346
합계			261506

표 2 2014년도의 현대소설 코퍼스 수록 작품과 어휘수

이와는 별도로 몇몇 소설 작품에 대해서는 어휘 분석까지 하고 있다. 현대소설에 대한 어휘 분석이 완성되면, 이 데이터 역시 한국 신소설 어휘사전 사이트와 통합하여 문학어 어휘사전으로 서비스할 계획이다.

2.5. 각종 문학 코퍼스의 제작

연구과제로 수행한 문학 코퍼스의 구축 외에도 필자는 여러 기회를 활용하여 문학 코퍼스를 확장해 나가고 있다. 그 주된 기회는 필자가 지도하고 있는 석박사 과정 학생

들이 논문을 작성할 때다. 필자는 학생들에게 작품론 차원에서 학위논문을 작성할 경우에 가급적 어휘 빈도를 통한 계량적 접근 방식을 택하도록 권장하고 있다. 특히 현대시인을 주제로 할 경우 기존의 현대시 코퍼스에 포함된 시인의 텍스트를 바탕으로 하고, 1950년 이후에 출판된 시집의 시 텍스트까지 포함하여 시인별 전집 코퍼스로 확장하기도 한다. 그 대표적인 것 몇 가지를 소개하면 다음과 같다.

구분	작가	규모(어절)	형태	가공정보	DB 처리	기타
의미분석용 시 코퍼스	김소월	7,948	동음이의어 및 다의어 분석말뭉치	어휘용례색인	DB수록	어휘의미분류처리
	정지용	7,651				
	조지훈	20,574				
	김조규	42,250				
	김현승	25,738				
	신석정	36,763				
개별 시 코퍼스	노천명	11,000	원시말뭉치	행별용례	DB 수록	
	김종삼	2,500	원시말뭉치	어절용례	DB 수록	
	김수영	22,000	원시말뭉치	원문입력		
어휘비교용 시 코퍼스	김억 창작시	11,000	원시말뭉치	행별용례	DB 수록	
	김억 번역시	2,500	원시말뭉치	어절용례	DB 수록	
	서지마(中)	31,459	기본형분석 말뭉치	어휘용례색인	DB 수록	한-중 병렬 말뭉치
	이시카와 다쿠보쿠(日)	18,936	기본형분석 말뭉치	어휘용례색인	DB 수록	한-일 병렬 말뭉치
수필 코퍼스	김학철	25,736	기본형분석 말뭉치	어휘용례색인	DB 수록	

	이상	18,357	기본형분석 말뭉치	어휘용례색 인	DB 수록	
구비문학 코퍼스[8]	용인설화	67,495	기본형분석 말뭉치	어휘용례색 인	DB 수록	진행중

표 3 김병선 교수의 문학 코퍼스 구성

이와 같은 문학작품 텍스트에 대한 분석용 코퍼스 외에 필자는 워드프로세서 문서 형식으로 작성된 문학작품 본문 파일도 다수 확보하고 있다. 향후 적절한 기회에 문학 코퍼스로 확장하여 수용할 예정이다.

텍스트 코퍼스 외에 다음과 같은 목록 정보도 필자의 문학 데이터베이스에 포함되어 있다. 이러한 목록 자료는 문학 코퍼스의 수집 계획을 세우거나 어휘 용례색인의 출처 정보를 표시하는 것 등에 연계하여 활용하고 있다.

(1) 현대시 총목록 (20,078종)

(2) 현대시집 총목록 (9,768종)

(3)『개벽』,『조광』 문학기사 총목록 (26,06종) (일부 본문 포함)

(4) 신소설 총목록 (304종)

(5) 현대소설 총목록 (약 1,000편)

[8] 구비문학 코퍼스는『한국구비문학대계』의 설화 자료를 수록한 것이다. 시범적으로 용인 지역의 설화를 코퍼스로 전환하여 보았다.

3. 문학 코퍼스의 구성과 구조

필자가 구축하고 있는 문학 코퍼스의 형식은 기존의 언어적 코퍼스와는 많이 다르다. 그것은 어휘 색인의 생성을 통하여 그에 대한 계량적 연구를 전제로 구축하고 있기 때문이다. 이제 기초말뭉치로부터 데이터베이스 형식으로 전환되기까지의 과정을 중심으로 문학 코퍼스의 구성 내용과 그 구조에 대해서 설명하기로 한다.

3.1 기초말뭉치로부터 데이터베이스로 변환

일차적으로 기초말뭉치는 워드프로세서 파일(hwp 포맷)로 작성한다. 작품별로(소설의 경우), 시집별로(시의 경우) 파일을 작성하되 헤더 파일에는 저자와 제목 및 출처 정보(수록 문헌의 코드)만 표기하고, 작품 본문은 일반적인 출판 양식에 맞추어 작성한다. 문헌 출처에 대한 정보는 별도의 데이터베이스 테이블로 관리한다. 시작품의 경우는 행말에 '/'를 붙여서 하나의 연을 하나의 단락(paragraph)으로 변환한다. (이는 용례 색인에서 문맥의 범위를 가급적 넓히고자 하는 것이다.)

기초말뭉치는 유니코드 텍스트 파일로 저장하여, 필자가 만든 용례색인 생성 프로그램인 <똑똑새> 처리를 거쳐서 용례색인 파일로 변환한다. 굳이 <똑똑새>와 같은 제3의 프로그램이 아니라도, <한글>의 스크립트 매크로(script macro) 기능과 '찾아 바꾸기' 기능 등을 활용하여 자료를 정리하고 이를 바탕으로 MS Excel 등에서 용례색인을 생

성할 수도 있다. 이때 생성되는 용례색인은 KWOC 형식을 따른다. 이후 이 텍스트 형식의 용례색인 파일은 MS Excel의 시트(sheet)나 MS Access의 테이블(table)로 수입된다.

3.2 기본적인 용례색인 테이블의 구조

어휘 용례색인 파일은 다음과 같은 기본 정보의 필드로 구성된다.

(1) 활용형: 원래의 텍스트에서 띄어쓰기가 되어 있는 어절을 각각 독립시켜 이를 활용형 필드에 저장한다. 이때 문장부호는 제외된다.

(2) 문맥: 활용형 어절을 키워드로 하고, 의미를 파악할 수 있는 정도의 길이를 지닌 앞 문맥과 뒤 문맥을 저장한다. 이후 이러한 문맥은 해당 어휘의 용례로 제시된다.

(3) 출처: 작자, 작품의 제목, 수록 작품집 및 출판연도 등의 정보를 저장한다. 출처의 범위는 기초말뭉치에서 표시한 태그에 따라 달라질 수 있다.

3.3 기본형 밝히기와 품사 분석

용례색인이 생성된 이후에는 각 필드별로 활용형 어절을 키워드로 하여, 어휘의 기본형 정보를 분석하여, 새로운 필드로 저장한다.

(1) 기본형: 텍스트로부터 추출된 활용형 어휘(어절) 각각에 대해서 기본형을 확정해 준다. 기본형 또는 원형을 밝히는 작업(lemmatization)은 일반적으로 전산언어학에서는 형태소 분석기라는 프로그램을 통하고 있지만, 문학 텍스트의 경우(특히 옛한글이나 비표준적 표기가 많은 시작품이나 신소설의 경우)는 자동 분석 프로그램보다는 전

문가의 수작업으로 할 수밖에 없다.

처음 시도했던 현대시 코퍼스에 대해서는 60여만 개의 어휘에 대해 수작업으로 기본형을 추출했다. 이후 신소설과 현대소설 등의 기본형 추출 작업에는 기분석 목록(현대시 시어의 기본형 확정 작업을 통해 활용형과 그에 대응하는 기본형 후보를 나열한 목록)을 적용하는 방식을 적용하여 그 효율을 높였다.[9] 기본형 밝히기 작업이 확대되면서 기분석 목록도 확장되고 있으며, 차후 다양한 장르의 텍스트에 대한 분석 경험을 활용한 지능형 분석기로 발전시킬 계획이다.

기존의 형태소 분석 프로그램이 단순히 기본형 추출에 그치는 것이라면, 필자가 적용하는 방법은 한 걸음 더 나아가 동음이의어 및 다의어까지 분석하는 것으로서 그 결과물은 어휘 연구에 곧바로 활용할 수 있는 수준이다.[10]

사실 기본형 밝히기 작업은 아무리 도구의 도움을 받는다고는 하여도 어휘론적 지식과 더불어 집중적인 작업을 필요로 하는 것임에 틀림없다. 문학 텍스트에 대해 어휘 차원에서 접근하는 연구자라면 기분석된 어휘 목록만을 활용하는 것보다는 자기 자신이 직접 이와 같은 기본형 밝히기를 해 봄으로써 대상 텍스트에 대한 적응 능력을 키우는 것이 바람직하다고 본다.

기본형 정보는 다음과 같이 표시된다.

---

[9] 기분석 목록 참조를 통한 기본형 밝히기에 대해서는 다음 논문을 참조. 김병선, 「시어의 기본형을 찾아서 -목록 참조를 통한 지능적 기본형 추출 방안-」, 『문학 연구와 정보과학 학술회의 자료집』, 한국학중앙연구원 어문생활사연구소. 2009. pp.5-28.

[10] 기본형을 확정하는 과정에서 오표기나 오류 어휘로 보이는 것들에 대해 다시 원본비평적인 고찰을 했다. 여러 번 발표된 작품의 경우 작가의 생존시 최종적으로 발표한 것을 결정본으로 채택했다.

[어휘의 기본형][(원어 표기)][동음이의어 구분용 첨자][(다의어 구분용 부호)][품사기호]

① 어휘의 기본형: 기본형을 한글로 표기한다. (아라비아 숫자, 영문자 등도 한글 표기법에 따라 한글로 표기한다.) 하나의 어절을 하나의 어휘로 확정한다. 즉 조사나 접미사 및 어미 등은 별도의 어휘로 구분하지 않는다. 기본형 확정에서 문제가 되는 것은 합성어와 활용과 곡용에 의해 변성되는 경우다. 합성어의 경우는 사전에 표제어로 등재되지 않은 경우라도 합성의 결과 새로운 의미가 생기거나 의미의 변용이 이루어지거나 의미가 좁혀지는 것을 직관으로 판단하여 하나의 어휘로 취급한다. 품사의 변성이 일어나는 경우 특히 형용사 뒤에서 '-어지다'의 구성으로 되어 '어떤 상태로 됨'을 뜻하게 되므로 이를 동사로 취급한다. 서술성을 가지는 명사 어휘 뒤에서 '-되다'의 구성으로 되어 '피동'의 뜻을 더하는 경우에는 동사로 취급하고, 몇몇 명사, 어근, 부사 뒤에 붙어 형용사를 만드는 경우에는 형용사로 취급한다.

② 원어 표기: 외래어나 외국어 및 한자어 등의 어휘에만 적용하는 선택적인 항목으로서, 한자어의 경우는 음절 위치에 맞추어 표기하고 한글 부분은 음절 위치에 '_' 표시로 대치한다. (예: 정신없이(精神__), 그전(_前))

③ 동음이의어 구분용 첨자:『표준국어대사전』에 근거하여 그 사전의 첨자를 두 자리의 아라비아 숫자로 표기한다. 동음이의어가 없는 경우에는 '00'을 붙인다. 만일 사전

에 표제어로 등재되지 않은 어휘라면 90번 대의 번호를 붙인다. (예: 가리다01, 가리다02, 가리다03, 가리다04 등. '01'은 '보이거나 통하지 못하도록 막히다', '02'는 '보이거나 통하지 못하도록 막다', '03'의 대표적 의미는 '여럿 가운데서 하나를 구별하여 고르다', '04'는 '곡식이나 장작 따위의 단을 차곡차곡 쌓아 올려 더미를 짓다'의 의미이다.)

④ 다의어 구분용 부호: 하나의 어휘가 두 개 이상의 품사로 『표준사전』에 등록된 경우에 선택적으로 표기하되, 사전의 순서에 따라 ㉠, ㉡, ㉢ 등을 붙인다. 더러 같은 단계에서 다시 나누어야 할 필요가 있는 경우에는 ㉮, ㉯, ㉰ 등을 덧붙인다. (예: 말다03㉠, 말다03㉡. 앞의 것은 동사이고, 뒤의 것은 보조동사로서 품사가 다르다.

⑤ 품사 기호: 「21세기 세종계획」에서 정한 품사 기호를 두 자리의 알파벳 기호로 다소 수정한 기호를 붙인다.

명사(ng), 대명사(np), 수사(nr), 고유명사(nm), 의존명사(nb)

동사(vv), 형용사(va), 보조동사(vx), 보조형용사(vz)

부사(ma), 관형사(mm), 감탄사(ic)

접두사(xp), 접미사(xs), 어근(xr), 조사(jk)

(예: 가지01ng (나무나 풀의 원줄기에서 뻗어 나온 줄기), 가지04nb (사물을 그 성질이나 특징에 따라 종류별로 낱낱이 헤아리는 말), 버리다01㉠vv (가지거나 지니고 있을 필요가 없는 물건을 내던지거나 쏟거나 하다.), 버리다01㉡vx (앞말이 나타내는 행동이 이미 끝났음을 나타내는 말))

이와 같은 형식으로 표시되는 어휘의 기본형 정보는 다른 테이블의 정보에도 동일하게 적용하였다. 말하자면 이것을 문학 데이터베이스에서 하나의 어휘 표준 코드로서 작용하는 것이다.

3.4 기타 분석 정보의 부가

기타 분석 정보는 현재 코퍼스의 종류와 용도에 따라서 적용된 것도 있고 그렇지 않은 것도 있는데, 향후에는 모든 코퍼스에 적용할 예정이다.

(1) 관련어 정보: 지역어-표준어, 속어와 비어-표준어, 준말-본딧말, 동의어 등을 제시한다.

(2) 어절의 구성 정보: 기본형과 원형의 어휘 및 어미나 조사를 제시한다. 어미나 조사는 별도의 필드로 독립시켜 처리할 수 있다.

(3) 품사 정보: 기본형 표기에 품사 기호가 표시되지만 이를 따로 제시한다.

(4) 어휘의 구성 정보: 고유어, 한자어, 외래어(일본어, 영어, 기타 외국어 등), 외국어 등의 구성 정보를 표시한다.

(5) 어휘의 등급: 한국어 교육용으로 판단한 어휘의 등급을 제시한다. 그 기준은 국립국어원의 조사 결과에 따르되, 어휘 등급 목록에 포함되지 않는 어휘에 대해서는 D등급과 E등급을 부여한다.

(6) 의미 분류: 필자가 귀납적으로 개발한 일종의 어휘 시소러스를 적용한다. 대분류의 구분 기준은 품사이며, 그 하위에 중분류와 소분류를 둔다. '중분류-소분류' 형식으로 표기한다. 의미 분류 작업은 매우 까다로운 것이기 때문에 위 코퍼스 구성 내역에

서 '의미 분석용 시 코퍼스'에만 적용한 상태다.

(7) 연어 정보: 일반적으로 2 gram 형식으로 연어 정보를 제시한다. 여기에는 두 가지 형식이 있다. 첫째는 두 개의 활용형 어휘를 연달아 표시하는 것이고, 둘째는 기본형 어휘를 연달아 표시하는 것이다. 이 연어 정보는 어떤 저자의 말씨를 파악하는 색인으로 활용될 수도 있고, 품사 분석의 오류나 기본형 분석의 오류를 바로잡는 용도로도 활용된다.

(8) 뜻풀이: 별도의 테이블에 어휘에 대한 사전적 정의를 저장한다.

4. 문학 코퍼스를 활용한 연구 과제와 그 전망

이러한 대규모의 문학 텍스트 코퍼스 구축을 해 오는 동안, 필자는 실제 문학 코퍼스를 구축할 뿐 아니라 문학 연구용 코퍼스 구축의 방법론에 대해서도 연구를 거듭하였다. 그리고 그 과정에서 이 코퍼스를 통해서 특히 계량적인 접근을 할 때에 그 기초 자료가 되는 어휘들과 그 어휘들이 모여서 이루어진 문학작품의 무결성이 가장 중요하다는 것을 절실하게 느끼게 되었다. 만일 하나의 어휘를 잘못 분석한다면 그 이후 이루어지는 각종의 데이터 가공 작업과 통계 작업에 고스란히 그 오류가 반복 또는 확장될 수 있기 때문이다.

그런데 문학 코퍼스에 대한 정보처리의 작업이라는 것이 단순한 정리 작업에 머물지 않고 역으로 문학 텍스트에서 오류를 발견하는 데에도 크게 기여한다는 것을 발견하

게 되었다. 말하자면 원래의 문맥에서 통합적(syntagmatic) 질서로 존재하던 문학작품의 어휘들이 문학어 데이터베이스를 통해 계열적(paradigmatic) 질서로 재편되는 과정에서 원본 표기의 일관성 문제와 오류 문제 등을 쉽게 발견해 낼 수 있었다는 것이다. 따라서 자료에 대한 숙고는 문학 텍스트에 접근하는 모든 연구자의 일차적인 의무이지만, 그가 다루는 텍스트의 범위가 매우 넓거나 분량이 아주 많은 경우에는 전산정보처리의 도움을 받아야 한다는 경험을 한 것이다.

필자의 문학 코퍼스는 일반적인 언어적 코퍼스와는 달리 헤더에 필요한 정보를 작품에 관련되는 것만으로 최소화하고 있다. 이는 문학 연구라는 특정 목적을 전제로 하기 때문이며, 그 결과 텍스트에 대한 가공 등에서 군더더기 없는 처리가 가능한 상황이다. 일반적인 언어적 코퍼스가 XML 스타일에 의한 마크업을 해놓은 것이라면 필자의 문학 코퍼스는 데이터베이스 형식으로 전환되어 있다. 그러나 만일 XML 문서 분석 프로그램을 응용해야 할 경우라면 필자의 코퍼스도 XML 형식으로 전환될 수 있다. 하지만 별도의 전용 프로그램이 아니라 MS Access나 MS Excel 등과 같은 범용 프로그램에서 데이터의 가공과 검색 그리고 상당한 정도의 통계 처리가 가능하다는 것은 일반적 연구자들에게는 매력적인 일이라 아니할 수 없다. 특히 MS Office 프로그램에서 지원하는 쿼리 언어(query language)를 통해 자료에 대해 지능적인 접근을 할 수 있기 때문에 고급의 정보처리를 원하는 연구자들에게도 매우 유익한 형식이 된다고 할 수 있다.

이제 이러한 문학 코퍼스에 접근하기 원하는 연구자를 위해 몇 가지 실제 연구 사례

를 소개하고, 앞으로의 확장에 대해서 얘기하겠다.

첫째는 문학 어휘에 대한 계량적 접근이다. 계량적 접근은 텍스트에 나타나는 어휘적 현상에 대한 객관적이고 수치적인 표현으로 나타내는 것을 기본으로 한다. 빈도 정보뿐만 아니라 상대 빈도나 빈도의 순위도 표현해 낼 수 있어야 한다.
문학 텍스트의 각 어휘가 용례색인 형태로 데이터베이스의 레코드로 전환되어 있으며, 여러 가지 부가적 정보가 데이터베이스의 필드로 저장되어 있으므로, 각 항목에 대해서 개별적인 통계도 가능하고, 두 개 이상의 필드 항목에 대해서도 통계작업이 가능하다. 이를테면, 여러 명의 작가를 대상으로 그들의 품사 사용 양상에 대해서도 간단한 쿼리 언어(Access의 경우)나 피봇 테이블 작성(Excel의 경우)을 통해서 쉽게 통계 처리를 할 수 있다. 나아가서는 더 진보된 통계 기능(예를 들면 평균치, 중간치, 최소치, 최대치 및 표준편차의 추출 등)도 간단한 함수를 적용하여 처리해 낼 수 있다. 문학 연구를 위한 수준이라면 전문적인 통계 패키지(SPSS 등)를 사용하지 않더라도 Excel의 내부 함수를 통해 웬만한 정도는 처리할 수 있을 것이다. 또한 통계 처리의 결과는 시각화(visualization) 도구를 통하여 적절한 그래프로 제시될 수도 있다.

둘째는 분석용 정보를 활용한 접근 방법이다. 언어학자들은 이러한 문학 코퍼스에서 어휘의 구성 분석과 같은 정보를 활용하여 어휘론적인 접근을 할 수도 있을 것이며, 특정 문법적 요소가 사용되는 환경도 찾아낼 수 있을 것이다. 또한 언어교육의 입장에서는 어휘의 등급에 대한 정보 분석도 매우 유용하게 활용될 수 있다. 각 작품별로 어

휘 등급표를 적용하여 작품의 이독성(readability)에 대한 객관적 평가가 가능하며, 이를 통해서 문학교육용 문학 텍스트를 선정한다든지, 난이도에 따라 교육 수준을 판단할 수 있는 좋은 자료로 활용할 수 있다.[11]

의미 분류 정보도 매우 유용하게 활용될 수 있다. 여섯 시인의 텍스트를 포함하고 있는 '의미 분석용 시 코퍼스'에는 모두 140,807개의 어휘가 포함되어 있으며 이를 모두 300 여개의 분류항목으로 구분하여 보았다. 이를 통해 이 코퍼스에 포함된 6명의 시인들의 시적 제재 혹은 의미 영역 등에 대한 연구가 가능하게 되었다고 할 수 있다. 이 방법은 특히 시인간의 비교 연구에 큰 도움이 될 수 있는데, 6 시인 중에서 신석정의 식물 시어가 압도적으로 많은 것을 확인할 수 있었다.

셋째는 이와 같은 분석 정보를 활용하여 보다 심도 있는 응용 연구를 할 수 있다. 특히 기대되는 것은 문체론적 연구(stylistic research)이다. 한 개인이나 한 시대 및 한 언어공동체의 문체는 통계적인 관점에서 기술될 수 있다. 연구 대상이 되는 텍스트에서 지배적인 어휘와 언어적 특성은 무엇인지, 비교 대상이 되는 텍스트와는 어떤 점에서 개별적인지를 기술하기 위해서는 위와 같은 계량적 정보가 필수적이다. 아직 계량적 접근을 통한 텍스트 분석 기술이 충분히 연구되지 않은 상태이지만, 이러한 연구가 계속된다면 틀림없이 좋은 결과가 있을 것으로 예상된다. 전문가의 직관적 문체 분석이 어떠한 계량적, 절차적 근거를 가지고 있는지를 탐색하는 것이 중요 과제일 것이

---

[11] 문학작품에 대해 어휘 등급을 적용하는 방안에 대해서는 다음 논문 참조. 김병선, 「한국 현대시의 언어 등급과 문학교육」, 『한국학연구논문집(1)』, 중국문화대학화강출판부, 2012.

다.

필자는 비교 대상이 되는 작가의 텍스트가 얼마나 유사한지를 판단할 수 있는 연구를 수행해 본 일이 있다.[12] 이런 텍스트 유사성 연구와 더불어 저자 판정 같은 분야도 앞으로 각광을 받을 수 있다고 본다. 특히 저작자를 알 수 없거나 저작성이 애매한 신소설 텍스트에 대해서 이러한 연구가 수행될 필요가 있다.

한편 2014년부터 시작하고 있는 한국현대소설 코퍼스 구축에 있어서는 문학적 특성을 반영한 구조를 모색하고 있으며(예를 들어 대화문과 지문의 구분, 대화자와 서술자에 관현 정보 태그의 부착 등), 이러한 자료를 통한 연구도 문학 연구의 지평을 확장할 수 있으리라고 본다.

하여튼 문학 텍스트에 대한 계량적 연구가 보다 폭넓게 진행되고, 이 분야의 연구 공동체가 확장될수록 그를 통해 얻는 결과도 풍부해지리라 기대해 본다. 필자는 이를 위해서 필자의 문학 코퍼스를 연구자들에게 제공하고 있으며, 이러한 연구 자료를 다루는 기법과 그 해석에 대한 지침서를 집필하고 있다. 만일 필자의 문학 코퍼스에서 어휘의 용례와 빈도 정보를 얻기 원한다면 일차적으로 누리미디어(주)의『한국현대시어 용례사전』데이터베이스와 한국문화사의『한국현대시어 빈도사전』, 그리고 한국학학술정보관의『신소설 어휘사전』을 활용하기 바라고, 보다 더 진전된 혹은 가공된 정보를 원한다면 다음과 같은 점에 유의하여 신청하기 바란다.

---

[12] 계량적 문체론에 대해서는 다음 논문 참조. 김병선(2006),「현대시인의 문체적 지문을 찾아서」,『국어국문학』제 143호, 국어국문학회. pp.153- 188. 김병선(2007),「시적 유사성 탐구 방안 연구」,『조선-한국학 국제학술대회 발표논문집』, 중국 연변대 아세아연구소.

(1) 연구의 대상과 분석의 구체적 내용 및 그 용도를 밝혀서 신청한다.

(2) 연구 결과물에 필자의 문학 코퍼스를 활용했다는 사사표기를 한다.

(3) 필자로부터 받은 자료의 무결성을 점검하여 이를 필자에게 통지한다.

(4) 문학 코퍼스의 확장을 위하여 새로운 텍스트를 필자에게 제공한다.