한•중 양국 현대문학의 실증적 연구

-서지마 시 텍스트 처리 사례를 중심으로-

요위위(姚委委)[1]

1. 문제제기

한•중 양국은 지리적으로 가깝고, 오래 전부터 문학을 포함한 문화 전반의 교류가 활발하게 이루어져 왔다. 이에 따라 그 동안 두 나라의 문학비교에 관한 연구는 많이 누적되어 왔다. 그러나 연구의 대부분은 고전문학을 중심으로 한 한•중 문학 간의 직•간접적인 영향관계를 탐구한 것들이었다.

고전문학의 비교연구가 활발한 데 비해 한•중 근대문학의 비교 연구는 보다 지연된 상황이다. 이것은 현대에 들어와서 한•중 양국의 문학 간에 직•간접적인 영향관계가 약해지고 수평적 비교 연구가 객관성이 부족한 지적을 받고 있다는 점과 무관하지 않다.

실증적 연구, 특히 계량적인 방법을 활용하는 것은 문학 비교 연구의 객관성을 높이는 데에 도움이 된다. 실증적 데이터를 가지고 비교대상의 어휘의 공통점과 차이점을 추출하고 그들이 어떤 어휘를 어떻게 구사하였는지를 밝히는 것은 비교 대상의 특징을 설명하는 데에 객관적인 ‘증거’를 제시할 수 있기 때문이다. 뿐만 아니라 문학 텍스트 어휘를 계량적인 방법으로 정리한 것은 한•중 양국 문체적 문학 흐름을 파악하고 한•중 양국 문학 언어 디지털 아카이브를 구축하는 데에 기여를 할 수 있다.

계량적인 실증적 연구에 있어 한국은 선도적 입장에 서 있다. 특히 한국학중앙연구원 김병선 교수를 비롯한 연구자들은 지난 10 년 동안 한국 현대시 데이터베이스를 구축하여 한국현대시 코퍼스(Korean Modern Poetry Corpus: 약칭 KoPoCo)를 만들고 『한국현대시어 용례사전』(2007)과 『한국현대시어 빈도사전』(2007)을 출판하였다. 특히 김병선 교수는 제 5 회 세계한국학 대회의 발표 논문 「어휘 비교를 통한 한중일 근대문학 연구 방안(2011)」에서 한중일 어휘비교의 비교 방법론을 제시한 바가 있다. 그의 방법론은 한중일 어휘 추출, 번역문제와 처리 등에 대한 요령적인 내용을 포함하고

[1] 中国河北大学外语学院教师, 한국호남대학교 조교수.

있다. 이와 같이 한국에서 현대문학에 관한 실증적 연구 성과물들은 한•중 양국의 현대문학 비교연구에 튼튼한 힘이 되어 준다.

본고에서는 서지마(徐志摩)[2] 원문 텍스트의 처리 사례를 중심으로 김병선 교수가 제시한 한중일 어휘비교 방법론의 큰 틀 아래서 실제적으로 중국어로 된 텍스트를 처리할 때 겪는 원전비평 문제, 원문텍스트의 띄어쓰기와 어휘 번역 등 문제에 대한 해결책과 해결 기준을 구체적으로 논의해 보고자 한다. 이로써 중국어 시 텍스트를 처리하는 방법을 구체화시키고 한•중 현대시 코퍼스를 만드는 데에 일조가 되고자 한다.

2. 한•중 양국 현대문학의 실증적 연구

가. 한국의 KoPoCo

KoPoCo 는 한국현대시 코퍼스의 약칭이며 1923 년부터 1950 년 사이에 출판된 창작 시집 소재의 현대시 작품과 작가들에 대한 정보를 데이터베이스화한 것이다. KoPoCo 에서는 시인 344 명, 한시나 번역시를 제외하고 자유시, 시조, 장편시 등 다양한 장르, 8201 편 현대시를 데이터화하였다.

KoPoCo 의 데이터를 추출하기 위해 시 원전의 확정, 기본형 추출, 동음이의어분석, 다의어 분석, 내면시어 분석, 한자 및 외래어 표기 분석, 품사 분석[3]등 과정을 거쳤다[4]. KoPoCo 데이터는 이미 각 시인의 시어가 추출되고 계량적인 처리 추출하고 계량적인

---

[2] 서지마(徐志摩), 중국 근대 시 문학사에서 유명한 시인이다. 그는 신월파(新月派)를 주도한 인물이며『지마의 시(志摩的詩)』·『비랭취의 일야(翡冷翠的一夜) 』·『맹호(猛虎) 』·『운유(雲遊) 』등 4 개의 시집,『낙엽(落葉) 』·『자부(自剖)) 』·『파리의 기억조각(巴黎的鱗爪) 』·『추(秋) 』·『서지마 일기(徐志摩日記) 』·『애미소찰(愛眉小札) 』등 6 권의 산문집과『와제해(渦堤孩) 』·『죽음의 성(死城) 』등 번역소설집이 있다.

[3] KoPoCo 에 수록한 시집에 대해 각자 원전비평, 기본형 추출, 어휘분석 처리 등 과정을 거쳤다. KoPoCo 에서 한국어 텍스트를 처리하는 방법을 자세하게 나와 있다. 이 논문은 주로 KoPoCo 데이터에 맞춰 중국어 텍스트를 처리하는 방법을 다루는 내용이라 한국시집과 한국시 데이터 처리 방법에 대해 자세히 언급하지 않았다. 중국어로 된 텍스트를 처리하는 방법은 KoPoCo 에서 한국어로 된 텍스트를 처리하는 방법과 <어휘 비교를 통한 한·중·일 근대문학 연구 방안>에서 비교연구 방법론을 참고하고 정리하였다.

[4] 김병선·조창환·배희숙·장노현 엮음, 『한국 현대시어 빈도 사전』, 한국문화사, 2007. p.1~8.참고

처리는 해 놓은 상태이다. 즉, KoPoCo 는 실증적인 한·중 문학 비교에 신빙성 있는 한국 측의 데이터를 제공해 주는 셈이다.

온전한 실증적인 한·중 문학 비교를 하기 위해 중국 측 텍스트를 KoPoCo 데이터와 같은 수준에 처리해야 한다. 또한 KoPoCo 의 성과물을 최대한 활용하기에 중국어로 되어 있는 텍스트를 한국어로 번역한 것이 유리하다. 이 과정에서 원전비평, 텍스트 구성, 띄어쓰기, 번역, 번역어휘 처리 등 작업이 이루어져야 한다.

나. 중국어 텍스트를 처리하는 방법

(1) 원전비평연구와 텍스트 구성의 원칙

(가) 원전비평연구

원전비평연구, 즉 신빙성 있는 원전을 확정하는 것은 정확한 데이터를 확정하기에 우선되어야 할 작업이다. 작가나 시인이 사후에 편찬된 문집은 물론이고 그들이 살아 있을 때 편찬된 문집이라도 오류가 있을 가능성을 배제할 수 없기 때문이다. 원전 확인과 교열 작업은 계량적인 연구 방법을 활용하기 위한 기초이다.

김소월과 서지마의 시문학을 비교하기 위해 서지마 시 텍스트의 원전 작업을 선행해야 한다. 서지마 시 원전은 중국의 역사의 문제로 찾기가 힘들지만 다행히 개혁개방 이후로 서지마 연구자들에 의해 편집 출판된 전집들이 있다.

	출판연도	판본	편집인	수록시수	수록한 특별작	판본의 특성
	1976	대만판	오복출판사	104	없음	중국어 번체를 채택했다. 문장부호가 현재의 것과 상당히 다르다. 문장이 가로쓰기로 되어 있다. 누락된 시가 많다.
	1983	상해판	상무인서관	104	〈중추월(仲秋月)〉	중국어 번체를 채택했다. 문장부호가 현재의 것과 상당히 다르다. 문장이 가로쓰기로 되어 있다. 대만판보다는 수록한 시가 많지만 여전히 빠뜨린 작품도

						많다.
3	1991	광서판	조하추	176		중국어 간체를 채택했다. 문장 부호를 주로 현재 통용되고 있는 것으로 고쳤다. 그러나 시 편집에 있어 많은 실수를 범했다. 수록한 시는 상해판보다 많다.
4	2005	천진판	한석산	184	〈하수시(賀壽詩)〉	중국어 간체를 채택했다. 문장 부호를 현재 중국 대륙에서 통용되는 것에 맞추었다. 편집에 있어 엄격한 태도를 취했다. 수록한 시로 볼 때 가장 완비된 상태의 판본이다.

위의 4 개의 판본을 살펴보면 □서지마 전집□은 시간이 지나면서 더 완비되어 가는 추세에 있다. 그 중 천진판은 보다 엄격한 편집 태도로 전집을 편집한 사실을 알 수 있다. 천진판에서는 186 수의 시를 서지마의 창작시로 취급해서 정리했다. 그러나 그 중 하나인「백화로 풀이하여 쓴 사 12 수(譯寫白話詞 12 首)」는 번안 성분이 들어가 있다. 또, 〈‘出其東門’을 백화로 풀어쓰다(“出其東門”白話寫意)〉은 중국 고문을 백화문으로 다시 해석한 것이다. 따라서 이 두 작품을 순수 창작시로 취급하는 것은 적당하지 않을 듯하다. 그렇다면 천진판에서는 모두 184 수의 창작시를 수록한 셈이다. 이 판본은 간체자로 되어 있고 문장 부호를 현대 중국에서 통용되는 것을 사용했으며 가로로 된 편집방식을 취했다. 또한 천진판에는 다른 판본에 수록되지 않았던 〈장수를 축하하는 시(賀壽詩)〉가 포함되어 있다. 천진판은 수록된 시 수가 가장 많으며 가장 엄격한 편집 태도를 보이는 판본이라 위의 네 가지 판본 중에서 가장 신빙성이 있다. 따라서 본고의 연구대상은 천진판에 수록된 시 184 수와 〈중추월(仲秋月)〉을 합친 총 185 수의 창작시로 확정한다.

천진판은 비교적 신뢰도가 높은 판본이긴 하지만, 소위 통용자라고 할 만한 글자들을 특별한 기준이 없이 사용하고 있는 문제점이 있다. 종이 책으로 읽을 때에는 독자들이 그것이 통용자라는 것을 알고 있기 때문에 큰 문제가 아닐 수 있지만, 컴퓨터 데이터 처리에 있어서는 서로 다른 글자로 인식하게 된다. 뿐만 아니라 온전한 결정본을 만들기

위해서는 통용자 처리에 대한 기준을 정해서 텍스트를 확정할 필요가 있다. 다음은 이 연구에서 밝힌 천진판의 문제점과 그에 대한 교열 원칙 및 한 사례이다.

> 重来此地，再检起诗针诗线〈康桥再会罢〉
> 检满一衣兜的贝壳-〈不再是我的乖乖〉
> 我检起一枝肥圆的芦梗-〈西伯利亚道中忆西湖秋雪庵芦色作歌 〉

‘检’과 ‘捡’: 다음 예문에서 ‘检’이 있는 자리에는 모두 ‘捡’이라는 동사가 들어가야 한다. 원작에는 ‘检’으로 되어 있는데, 독자들이 읽을 때 충분히 이해할 수 있기 때문에 천진판 편집자가 굳이 ‘捡’으로 고치지 않은 듯하다. 그러나 본고에서는 다른 작품에 쓴 ‘检’과 구별하기 위해 두 군데의 ‘检’을 ‘捡’으로 바꾸었다.

이처럼 선본의 오류, 통용자, 문장부호 등 문제에 대해 교열 작업이 이루어져야 신빙성 있는 데이터를 구성하기 기대할 수 있다.

서지마 시 텍스트 원전비평을 하면서 원전 확인의 원칙과 교열 작업의 범위는 다음과 같이 정리하였다.

1) 하나의 작품이 두 번 이상 수록된 경우에는 작가나 시인의 생존 기간에 발표된 가장 마지막 발표 작품을 택한다.

2) 작가나 시인 개인의 작품집에 수록된 작품과 여러 작가와 시인의 작품이 함께 수록된 작품집 중에서는 개인 작품집을 우선한다. 개작한 작품이 있을 경우에 원작품과 개작품은 모두 수록한다.[5]

3) 작가나 시인 생전 판본은 구하지 못한 경우 각 종류의 개인 작품집보다 참고할 만한 전집을 우선한다. 단 원문 텍스트 교열 작업할 때 개인 작품집을 참고해서 보완해야 한다.

4) 내용면에서 원본 자료집에 수록된 작품들은 내용대로 수록하는 것을 원칙으로 한다. 다만 번체 표기로 된 원본 자료집을 1986 년에 공포한 『簡化字總表(간화자 총표)』에 의해 중국어 간체로 바꿔야 한다.

5) 입력할 원본을 확정한 후 원본에 수록한 모든 작품에 대한 원전 비평을 거쳐야 한다. 원본 확정과 교열 과정에 원본의 내용을 존중하는 것을 원칙으로 하지만 내용이나 표기상 명백하게 오류라고 판단될 경우에는 교정 처리를 해야 한다. 원전 비평은 원본의 제목 표시, 문장의 행과 단락 변화, 문자오류, 통용자 처리, 문장부호의 변화 등을

[5] 김병선·조창환·배희숙·장노현 엮음, 앞의 책, p.5.참고.

고찰대상으로 삼는다. 이것은 신빙성 있는 원문 텍스트를 확정하기 위해 실행하는 작업이다.

(2) 텍스트 구성의 요소

KoPoCo 에 수록하는 한국시 원본 텍스트는 모두 〈한글〉(한글과컴퓨터사)에 입력되어 있다. 그러나 〈한글〉은 한자 입력, 특히 간체자 입력 기능은 수월하지 못해 중국어로 된 작품을 입력할 때 MS Word 중국어판을 활용하는 것이 좋다. MS Word 의 파일 형식은 〈한글〉과 다르지만 모두 유니코드 기반의 워드프로세서이기 때문에 최종적으로 UTF-8 텍스트 파일로 저장하여 코드체계를 맞출 수 있다.[6]

투명한 문서를 생산하기 위해 중국어로 된 작품을 입력할 때 작품 텍스트를 헤더 부분과 본고 부분으로 구분하여야 한다. 헤더에는 시인, 제목, 시집, 장르, 기타 등을 마크업 처리하고, 시 본문에는 행과 연 구분을 분명히 처리해야 한다.[7]

입력 형식의 예를 들면 다음과 같다.

〈시인〉徐志摩
〈제목〉偶然
〈시집〉翡冷翠的一夜
〈장르〉창작시
〈기타〉
我是天空里的一片云,
偶尔投影在你的波心—
…

(3) 띄어쓰기 및 중국어 용례색인 만들기

(가) 띄어쓰기

[6] 김병선, 앞의 논문. p.6.
[7] 김병선, 앞의 논문. p.7.

중국어는 한국어와 같이 어절별로 띄어쓰기 있어 각 어절에서 기본형을 추출하면 되는 문자가 아니다. 계량적인 연구방법을 적용하기 전에 문장 단위로 되어 있는 중국어를 띄어쓰기를 해서 어휘를 추출해야 한다. 그러나 중국어 어휘는 띄어쓰기가 쉽지 않다. 중국어 문장을 띄어쓰기 할 때 다음과 같은 원칙을 적용하기로 한다.

1) 최대한 어휘별로 분석한다. 중국어에서는 크게 실사(實詞)와 허사(虛詞)로 나누어져 있다. 실사는 명사, 동사, 형용사, 수사, 양사, 대사, 부사, 의성의태어 등을 포함하고, 허사는 개사, 접속사, 조사, 감탄사 등을 포함한다. 최대한 어휘별로 분석한다는 것은 명사가 명사대로, 동사가 동사대로, 형용사가 형용사대로 띄어쓰기를 한다는 것을 의미한다.

2) 개사, 접속사, 감탄사는 중국어의 허사에 해당한다. 즉 중국어에서 단독적으로 문장성분이 될 수 없는 어휘 종류다. 그러나 띄어쓰기 할 때는 개사 '把, 从, 向, 朝, 为, 为了, 往' 등, 접속사 '和, 及, 或者, 因为……所以, 不但……而且', 감탄사 '喂, 哟, 嗨, 哼, 哦, 哎呀' 등을 앞 뒤 연결하는 내용과 띄어쓰기 하기로 한다.

3) 조사인 '着, 了, 过, 吗, 呢, 吧'는 앞 어휘와 '-'로 연결한다.

4) 방위사(方位詞)는 방향 또는 위치를 표시하는 명사로써 앞의 명사와 띄어쓰기 한다. 예) 街道 上, 学校 里 등.

5) 중국어 합성어와 파생어는 한 단어로 간주한다. 예)
电灯, 记住, 房间, 面熟, 老师, 老虎, 仅仅 등

6) 중국어 이합사는 기본적으로 한 단어로 본다. 그러나 상황에 따라 띄어쓰기를 필요할 때 띄어쓰기 해야 한다. 이것은 이합사는 대로 나눠서 쓰고 대로 합쳐서 쓸 수 있는 어휘이며 중국어에서 특별한 어휘 종류인 것과 관련이 있다. 예를 들어 道歉, 洗澡, 吃饭, 画图 등이 그것이다. 제 5 판 『현대한어대사전(现代汉语词典)』에서 3400 개 정도의 이합사를 수록하였다. 이를 띄어쓰기 할 때는 한국어 번역어를 참고해서 처리해야 한다. 예를 들어 '道歉, 洗澡'는 한국어에서 대응하는 단어 있어 한 단어로 간주하지만 '吃饭, 画图 ' 경우에는 띄어쓰는 것은 마당하다.

7) 'ABC'형 어휘를 처리할 때 만약 'AB'가 있고 'BC'도 있으며 'ABC'도 있을 경우는 뜻을 보고 띄어쓰기를 해야 한다. 예) 半空 中. 만약 'AB'가 있고 'BC'도 있는데 'ABC'가 없을 경우에도 실제 내용을 보고 어떻게 띄어쓰기 할지를 결정한다. 예) 晚钟 声, 明月 下, 明月 夜

8) 개인적인 시어나 일반적인 합성어와 파생어에 속하지 않으며 중국과 한국 사전에 등록되어 있지 않지만 한 단어로 봐야 하는 단어는 띄어쓰기를 하지 않는다. 예) 林鸟, 서지마 시 원문에서 '苏醒的林鸟'라는 문장이 있다. 여기 '林鸟'를 띄어쓰기를 하면 '林'과 '鸟' 모두 중심어가 되어 버리는 오류가 생긴다. 여기서 중심어는 '鸟'이기 때문에 중국어 사전이나 한국어 사전에 모두 '林鸟'라는 어휘는 없지만 띄어쓰기를 하지 않는다. 이와 비슷한 예는 诗针, 诗线 등이 있다.

9) 비록 한국어에서 대응하는 한자어가 있다 하더라도 뜻이 다를 때는 원문을 띄어쓰기를 해야 한다. 예) 修 好, 한국어에 비록 '修好(두 나라가 사이 좋게 지냄)'라는 한자어가 있기는 하지만 원문에서는 '선행을 베풀다'라는 뜻이다. 따라서 이러한 경우는 띄어쓰기를 해야 한다.

10) 사자성어, 속어, 관용어 등 습관화 된 중국어 어휘는 그대로 유지한다. 예) 不可名状, 鳞次栉比, 无所不包, 半死不活, 有意无意, 无依无伴, 心心相印, 悲天悯人, 一丝半缕, 猫儿哭耗子,

11) 숙어나 사투리는 그대로 유지한다. 예) 可不是, 赶明儿, 妈妈呀, 格拉, 是勿是, 不妨事, 恨不能, 喔唷, 听说, 有人说, 见个儿, 恨不能

12) 외래어나 외국어는 그대로 유지한다. 예) "眸冷骨累"(melancholy), 印曼桀乃欣, 烟士披里纯

13) 의성・의태어는 그대로 유지한다. 예) 浩唉!, 割麦插禾

중국어 문장을 띄어쓰기 작업을 할 때는 최대한 어휘별로 분석해야 하지만 그렇게 할 수 없는 어휘들은 종류별로 처리해야 한다. 또한 투명한 문서를 만들기 위해 띄어쓰기 작업할 때는 일관성을 유지해야 한다.

(나) 중국어 용례색인 만들기

띄어쓰기 작업을 실행하고 어휘 분리를 한 다음에는 원문 작품 일차적으로 중국어로 된 용례색인(concordance)을 만든다. 입력된 텍스트는 중문판 워드프로세서에서 텍스트(*.txt)로 저장한 다음, <똑똑새>[8] 프로그램을 통해서 용례색인을 자동으로 생성한다. 또한 용례색인에서 충분한 문맥 정보를 보이기 위해 입력한 시 원고 텍스트에서 행 구분은 빗금(/)으로, 연 구분은 쌍빗금(//)으로 치환해 두었다. [9] 중국어로 된 용례색인에는 다음과 같은 내용이 포함되어 있다.

□ ID: 원문의 출현 순서를 나타내는 번호다.

□ 표제어: 기본형의 활용형 어휘를 제시한다.

□ 표기형: 원문에 쓰인 어절 키워드를 제시한다.

□ 조사: 앞 문장에서 활용하는 조사를 제시한다.

□ 번역어: 원문에 쓰인 어절 키워드의 번역어를 제시한다.

---

[8] 김병선 교수 제작한 어휘처리용 프로그램이다.

[9] 김병선, 앞에 논문, p.7. 참고.

□ 앞 문맥: 키워드의 앞쪽에 있는 문맥의 일부분이다. (똑똑새 프로그램에서 길이를 조절할 수 있다.)

□ 뒤 문맥: 키워드의 뒤쪽에 있는 문맥의 일부분이다. (똑똑새 프로그램에서 길이를 조절할 수 있다.)

□ 저자: 원문의 저자를 제시한다.

□ 제목: 작품의 제목을 제시한다.

□ 행 번호: 어절 키워드 원문에서 어느 행에 있는 것을 제시한다.

□ 출전: 작품의 출전 시집과 간행 연도 등을 제시한다.

□ 장르: 작품의 장르 유형을 제시한다.

□ 기타: 헤더 부분에 마크업 된 정보를 일괄 제시한다.[10]

(4) 번역과 번역어 처리

일반적으로 서로 문자체계가 다른 양국 문학 작품에 사용된 어휘를 비교할 때는 중간언어가 필요하다. 비교 대상이 되는 출발 문학 작품과 목표 문학 작품을 하나의 언어로 번역해야 서로 비교가 가능하기 때문이다. 중간언어는 자연어일 수도 있으나 인공어일 수도 있다. 한•중 양국 시어휘 비교할 때 KoPoCo 를 최대한 활용하면서 효과적으로 한국 현대시를 비교하기 위해 새로운 중간언어를 쓰는 것보다 중국어로 된 텍스트를 한국어로 번역하는 것이 유리하다.

서지마 중국어 시어 종수는 6606 종, 시어 개수 총 31,459 개 있고 번역 시어 종수는 5466 종이고, 시어 개수 총 31.459 개 있다. Excel[11]에서 서지마 번역시어의 형식을 보면 다음과 같다.

번호	어휘	개수	번역

[10] 구조상 중국어 어휘 색인의 틀은 한국어 어휘 색인과 비슷하다. 이 색인을 만드는 방법은 김병선 교수의 방법론을 따랐다.

[11] 어휘 번역은 MS Excel 이나 MS Access 에서 정리하는 것이 유리하다. 두 가지 프로그램의 기능을 활용해서 통계수치를 내거나 이중적인 번역을 확인할 때 편리하기 때문이다. 또한 어휘 번역할 때는 기계번역을 활용하는 것도 도움이 된다. 그러나 기계번역의 도구를 이용할 때는 번역 효율이 높은 도구를 선택해야 하고 번역된 어휘를 하나씩 하나씩 재검토 작업을 행해야 한다. 서지마 어휘를 번역했을 때 삼성전자의 <정음 Global>이라는 워드프로세서에 내장된 자동번역기(중-한)를 이용하였다.

1	感谢 01	1	감사하다 05vv
2	感谢 02	2	감사 08ng
3	诗	3	시 13ng
4	诗线	1	시 13ng
5	诗针	1	시 13ng
6	怪 01	6	이상하다 00va
7	怪 02	6	원망하다 01vv
8	怪 03	1	괴물 00ng
9	秦淮河	1	진회하 90nm
10	想象	3	상상 07ng
11	烟士披里纯	1	영감 03ng
… …	……	……	……

서지마 중국어 시어의 번역과 번역된 시어를 정리하면서 다음과 같은 원칙을 정했다. 이 원칙은 서지마 시어뿐만 아니라 KoPoCo 를 활용하는 한•중 문학 비교 연구에 일반 중국어 시어를 한국어 어휘로 번역할 때도 적용할 수 있다.

1) 어휘를 번역한다. 번역은 어휘번역과 문장번역으로 구별할 수 있다. 한·중 시어 비교 연구는 어휘를 중심으로 한 비교연구라 어휘 번역을 선택한다. 단 어휘의 품성이나 정확한 뜻을 파악하기 위해 앞 뒤 문장을 참고해야 한다.

2) 품사 종류를 구별해서 번역한다. 중국어에서 하나의 어휘가 두 가지 혹은 두 가지 이상의 품사 성격을 가지는 경우가 종종 있다. 예를 들면 '感谢', '爱' 등이 있다. 이러한 어휘를 번역할 때는 앞뒤문장을 참고해서 명사인지 동사인지를 구별해서 번역한다. 예를 들어 '감사하다 05vv', '감사 08ng'이 그것이다. 또한 중국어 원문 시어를 구별하기 위해 중국어 시어에도 번호를 붙여 준다. 예를 들면 '感谢 01', '感谢 02' 등이 그것이다.

3) 의미어 설정하고 이음동의어 또는 유사어를 번역한다. 어휘 비교에 있어서는 표준적인 어휘를 설정하고, 그것을 기준으로 삼아 비교해야 한다. [12]중국어 어휘를 한국어로 번역할 때는 이음동의어 또는 유사어 번역문제가 발생한다. KoPoCo 에서 김병선 교수님이 '의미어(semantic term)'라는 개념을 제시하신 바가 있다. 의미어는 각 이음동의어 또는 유사어의 표기형에 대한 표준적인 관련어(associate term)를 뜻한다. 예를 들면, '태양'과 '해', '햇님', '해님'에 대하여 '태양'을 의미어로 하는 것을 말한다. [13]중국어 시어를 한국어로 옮길 때 '의미어(semantic term)'라는 개념을 활용한다. 서지마는 같은 뜻을 표현할 때도 많은 표현하는 방법을 동원하였다. 예를 들어 '灵潮'나 '灵苗는 '灵感'의 다른 표현이다. 이러한 경우에 '灵感'라는 의미어를 설정해서 동일하게 번역한 것이다.

[12] 김병선, 앞의 논문. 2011.

[13] 김병선, 앞의 논문. p.9.

4) 표준어로 통일해서 사투리와 비표준적인 어법 시어를 번역한다. 예를 들면 '格拉', '吾', '俺家' 등이 모두 '我'의 사투리 표현인데 이를 번역할 때는 '我'로 통일해서 번역한 것을 말한다.

5) 원관념을 살려 비유어와 상징어를 번역한다.[14] 중국어에서 비유어와 상징어를 쓰는 것은 일반적이다. 특히 시나 소설 등 문학 작품에서 이러한 경우가 더욱 많다. 서지마 시어에서 '爱'를 상징하는 원관념 시어는 '明星', '大海', '玫瑰' 등이 있다. 이러한 경우에 시어의 원관념만 번역한다. 즉, '明星', '大海', '玫瑰' 를 '별', '바다', '장미'로 번역한 것이다.

6) 인칭대사 복수를 단수형으로 바꿔 번역한다. '我们(우리)'만 제외하고 '他们', '她们' 등이 '他', '她'로 바꿔 번역한다.

7) 사자성어와 숙어를 최대한 살려서 번역한다. 중국어 사자성어와 숙어를 번역할 때 한국어 대응하는 어휘가 있을 때는 대응하는 어휘로 번역하고, 없을 경우에는 한자어로 표시하고 90 번 번호를 붙여 준다. 예를 들어 '心心相印'를 '심심상인 00ng', '不可名状'를 '불가명상 90ng'로 번역한 것을 말한다.

8) 외래어를 표준어로 번역한다. 서지마 시어에서 '烟士披里纯'라는 외래어가 있는데 이는(inspiration) '영감'과 '인스피레이션'으로 번역 가능하지만 국어대사전에서 '영감'이라는 번역만 등록되어 있어 '영감'으로 번역한다.

9) 중국어 어휘 중에 마땅히 대치할 만한 한국어가 없는 경우에 일반적으로 원문 시어를 한자어로 표시하고 뒤에 '90'라는 번호를 부여 주고 구별한다. 예를 들면 '满家弄'을 '만가농 90'으로 번역한다.

10) 정확성과 통일성을 유의하여 번역한다. 중국어를 한국어로 번역할 때는 어휘 범위의 변화, 품사 성격 변화 등의 문제들이 있다. 번역할 때는 이러한 문제들을 주의하면서 번역해야 한다. 또한 신빙성 있는 통계결과를 얻으려면 번역의 통일성도 지켜야 한다. 즉, 같은 어휘는 같은 번역, 같은 상황은 같은 원칙으로 적용해야 한다는 뜻이다.

11) 번역된 어휘를 기본형으로 표기하고 품사 성격을 표시해 주는 기호와 번호를 붙여 준다. 예를 들어 감사하다 05vv, 시 13ng, 원망하다 01vv 등이 그것이다. 번역 어휘, 특히 기계 번역을 활용할 때 기본형으로 되어 있지 않은 경우가 많다. 이러한 경우에 어휘 하나 하나 기본형으로 바꿔 주어야 한다. 또한 정리된 기본형들이 '국립국어원 국어대사전'의 기준에 뜻 분류를 해야 한다. '감사하다 05vv'는 국립국어원 국어대사전에서 해당 어휘의 5 번 뜻이고, 'vv'는 품사 종류가 동사를 의미하는 것이다. 번역시어의 품사이름과 해당 기호는 다음과 같다.

[14] 김병선, 앞의 논문. p.9 참고.

품사이름	명사	동사	형용사	대명사	부사	조사	관형사	수사	의존명사	감탄사	고유명사	보조동사	어미	접두사	접미사	보조형용사	미확인	어근
기호	ng	vv	va	np	ma	jk	mm	nr	nb	ic	nm	vx	ed	xp	xs	vz	un	xr

표: 번역시어의 품사이름과 해당 기호[15]

다. 김소월과 서지마 시어 비교 연구

(1) 김소월과 서지마 시어 품사별 구성 상황

품사이름	기호	서지마(Z)		김소월(S)		KoPoCo	
		개수	비율	개수	비율	개수	비율
명사	ng	9920	31.959%	3,184	40.055%	250,928	41.009%
동사	vv	8150	26.256%	2,114	26.595%	149,209	24.385%
형용사	va	3346	10.780%	742	9.335%	58,808	9.611%
대명사	np	3185	10.261%	466	5.862%	32,928	5.381%
부사	ma	2427	7.819%	649	8.165%	45,454	7.429%
조사	jk	1247	4.017%	9	0.113%	1,836	0.300%
관형사	mm	1411	4.546%	278	3.497%	23,672	3.869%
수사	nr	71	0.229%	15	0.189%	1,748	0.286%
의존명사	nb	543	1.749%	177	2.227%	16,909	2.763%
감탄사	ic	335	1.079%	57	0.717%	4,425	0.723%
고유명사	nm	270	0.870%	74	0.931%	7,106	1.161%
보조동사	vx	59	0.190%	150	1.887%	15,534	2,539%
어미	ed	40	0.129%	0	0.000%	3	0.000%
접두사	xp	21	0.068%	0	0.000%	77	0.013%
접미사	xs	9	0.029%	1	0.013%	88	0.014%
보조형용사	vz	4	0.013%	28	0.352%	2,253	0.368%
미확인	un	2	0.006%	2	0.025%	482	0.079%
어근	xr	0	0.000%	3	0.038%	424	0.069%

[15] KoPoCo 어휘 품사이름을 표시하는 기호 체계와 동일하다.

총합계	31,040	100.000%	7,949	100.000%	7,949	100.000%

표: 서지마-김소월-KoPoCo 2011 시어의 품사별 빈도표

서지마와 김소월 작품의 시어를 품사별로 한국현대시 모집단과 비교해 보면 다음과 같다.

- 명사의 경우에 Z 와 S 는 큰 차이를 보인다. 김소월은 명사 사용 비율이 40%가 넘었지만, 서지마는 32% 밖에 미치지 못한다. 그러나 작품의 양을 고려한다면 사용하고 있는 명사의 종류는 Z 가 훨씬 많다.

- Z 의 동사 사용 비율은 S 의 동사 사용 비율과 거의 비슷하다.(26.256:26.595)

- Z 의 형용사 사용 비율은 S 의 형용사 사용 비율보다 약간 높지만 큰 차이는 없다.(10.780:9.335)

- Z 의 대명사 사용 비율은 S 의 것과 많이 다르다. Z 는 사용 비율이 10%를 넘어가는 것에 비해 S 는 6%를 넘지 못한다. 이것은 근대 한국어에서 대명사의 사용이 현재와 같이 다양하게 이루어지지 않았던 것과 관련이 있다. 또, 이는 한국어에서 대명사를 많이 생략하는 것에 비해 중국어에서는 대명사를 생략하지 않는 언어 특징을 지닌 것과도 무관하지 않다. 게다가 Z 의 시 평균 길이는 S 의 시 길이보다 4 배 정도 길다는 것을 전제로 한다면 두 시인의 작품은 대명사 사용 비율에서 큰 차이가 있음을 알 수 있다.

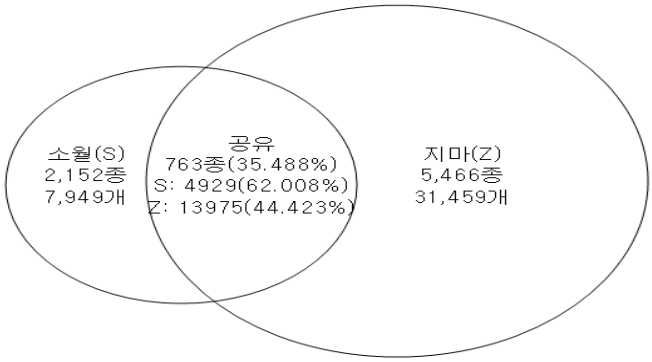
- Z 의 부사 사용 비율은 S 의 부사 사용 비율(7.819:8.165)보다 약간 낮다.

- 기타 조사, 어미, 접미사, 접두사 등 교착어에 대해서는 특별한 의미를 두기 힘들다. 사용 빈도가 높지 않을 뿐만 아니라 두 시인 사이의 차이도 그리 크지 않다.

(2) 김소월과 서지마 공유 시어 검토

공유 시어란 비교 대상이 되는 두 시인의 작품에 모두 나타나는 어휘를 말한다. 공유 시어를 시어의 출현 빈도와 더불어 검토하는 것은 작가나 시인의 작품의 유사성을 밝히는 하나의 유력한 방법이 될 수 있다.

공유시어를 추출할 때 처리된 Z(서지마)의 번역 시어 빈도와 S(김소월)의 시어를 적절한 컴퓨터 처리를 통해서 두 시인의 전체 시어(어종) 목록을 만들고, 시인별로 출현 빈도 필드에 쿼리를 통해서 각각의 빈도수를 추가한 작업을 먼저 이루어져야 한다. 또한 빈도에 공통적으로 1 이상의 값이 들어 있는 어휘를 두 시인의 공유시어로 설정하였다. 김소월과 서지마 공유시어를 도시화하면 다음과 같다.



전체 어종(type)의 값은 김소월(S) 2,152 종과 서지마(Z) 5,466 종의 합에서 공유되는 어종을 한 차례 제외하는 것으로 얻을 수 있다.(2,152+5,466-763) 집합으로 표시하면 S∪Z = 6,855 종이다.

전체 어휘(token)의 값은 김소월(S) 7,949 개와 서지마(Z) 31,459 개를 합하면 된다. 집합으로 표시하면 S∪Z = 39,408 개다.

공유 어종(type)의 값은 김소월과 서지마가 공유하고 있는 어종의 수를 계산하는 것으로, 집합으로 표시하면 S∩Z = 763 종이다.

공유 어휘(token)의 값은 김소월과 서지마가 공유하고 있는 어휘의 수를 계산하는 것으로, 소월은 4,929 개, 서지마는 13,975 개이다.

비공유 어종(type)의 값은 비교 대상과 공유하지 않는 어종의 수를 계산하는 것으로, 시인별 어종수에서 공유하고 있는 어종수(763 종)을 제외하면 된다. 김소월은 1,389 종(2,152 종-763 종), 서지마는 4,703 종(5,466 종-763 종)이다.

비공유 어휘(token)의 값은 비교 대상과 공유하지 않는 어휘의 수를 계산하는 것으로, 시인별 어휘수에서 각각 공유 어휘수를 제외하면 된다. 김소월은 3,020 개(7,949 개-4,929 개), 서지마는 17,484 개(31,459 개-13,975 개)이다.

이 계산의 결과를 표로 나타내고, 그 각각의 비율을 살피면 다음과 같다.

	어종			어휘		
	공유	비공유	계	공유	비공유	계
전체	763 종	6,092 종	6,855 종	18,904 개	20,504 개	39,408 개
	11.131%	88.869%	100%	47.970%	52.030%	100%

김소월	763 종	1,389 종	2,152 종	4,929 개	3,020 개	7,949 개
	35.455%	64.545%	100%	62.008%	37.992%	100%
서지마	763 종	4,703 종	5,466 종	13,975 개	17,484 개	31,459 개
	13.959%	86.041%	100%	44.4237%	55.577%	100%

- 공유 어종과 어휘의 비율이 높을수록 김소월과 서지마의 유사성도 커진다고 평가할 수 있다.

- 김소월과 서지마 시의 공유·비공유 어종 및 어휘의 비율을 살펴본 결과 전반적으로 공유 어종이 공유 어휘에 비해 그 비율이 매우 낮게 나타났다. 김소월의 경우 공유 어휘는 62%인 것에 비해 공유 어종은 36%에 미치지 못한다. 서지마는 그 편차가 더 커서 공유 어휘가 44% 이상인데 공유 어종은 14% 정도이다. 이러한 현상은 김소월과 서지마가 집중적으로 사용한 어휘들의 공유성이 높다는 것을 의미한다. 따라서 두 시인이 사용한 주요한 공유 시어를 추출해서 분석하는 것이 두 시인의 시 세계를 고찰하는 데에 필수적으로 요청된다.

김소월과 서지마가 사용한 중요 공유시어(significant shared word)를 선정하기 위해 두 가지 과정을 거쳤다. 하나는 해당 시어가 전체 시어에서 차지하는 비율을 측정하여 양 시인이 일정 수준 이상의 비율로 사용한 시어만을 선정하는 것이다. 아래의 표에서 S 의 비율은 S 의 빈도를 S 의 전체 시어(7949)로 나눈 백분율 값이고, Z 비율은 서지마의 경우다. 그리고 그 백분율 값 차이의 절대값을 비율차로 계산하였다.

그리고 다른 하나는 순위를 따지는 방법을 적용한 것이다. S 순위는 김소월이 쓴 전체 시어 중 S 빈도 값으로 순위를 매긴 것이고, Z 순위는 서지마의 경우다. 그리고 그 순위의 차이를 절대 값으로 표시한 것이 순위차 필드다. 본고에서는 김소월과 서지마가 쓴 상위 100 위 이내의 시어를 선택하기로 한다.

어휘	S 빈도	S 순위	S 비율	Z 빈도	Z 순위	Z 비율	비율차	순위차
나 03np	187	1	2.352%	1329	1	4.225%	1.873%	0
없다 01va	44	16	0.554%	161	18	0.512%	0.042%	2
어디 01np	21	61	0.264%	57	63	0.181%	0.083%	2
있다 01ㄴva	74	6	0.931%	290	10	0.922%	0.009%	4
그 01ㄴmm	61	9	0.931%	216	14	0.687%	0.244%	5
마음 01ng	39	24	0.491%	106	29	0.337%	0.154%	5

우리 03np	34	35	0.428%	83	40	0.264%	0.164%	5
다시 01ma	36	30	0.453%	112	24	0.360%	0.093%	6
하늘 01ng	39	25	0.491%	101	31	0.321%	0.170%	6
집 01ng	19	72	0.239%	47	81	0.149%	0.090%	9
알다 00vv	25	49	0.315%	85	38	0.270%	0.045%	11
이 05ㄴmm	41	20	0.516%	416	8	1.322%	0.806%	12
보다 01ㄱvv	35	33	0.441%	134	21	0.426%	0.015%	12
좋다 01va	21	62	0.264%	49	76	0.156%	0.108%	14
오다 01ㄱvv	104	2	1.308%	142	19	0.451%	0.857%	17
가다 01ㄱvv	99	3	1.245%	141	20	0.448%	0.797%	17
그러나 00ma	28	41	0.352%	128	22	0.407%	0.055%	19
꿈 01ng	40	22	0.503%	72	44	0.229%	0.274%	22
또 00ma	21	60	0.264%	106	30	0.337%	0.073%	30
죽다 01vv	38	26	0.478%	60	57	0.191%	0.287%	31
눈 01ng	15	98	0.188%	57	64	0.181%	0.007%	34
사랑 01ng	19	71	0.239%	91	35	0.289%	0.050%	36
바다 00ng	25	50	0.315%	46	87	0.146%	0.169%	37
세상 01ng	34	36	0.428%	50	74	0.159%	0.269%	38
구름 01ng	25	51	0.315%	45	90	0.143%	0.172%	39
바람 01ㄱng	36	31	0.453%	53	71	0.169%	0.284%	40
그 01ㄱnp	26	48	0.327%	417	7	1.326%	0.999%	41
사람 00ng	52	12	0.654%	59	58	0.188%	0.466%	46
때 01ng	78	5	0.981%	60	56	0.191%	0.790%	51
되다 01vv	49	13	0.616%	56	66	0.178%	0.438%	53
누구 00np	18	81	0.226%	112	25	0.356%	0.130%	56
밤 01ng	54	11	0.679%	52	72	0.165%	0.514%	61
아니다 00va	15	97	0.189%	108	27	0.343%	0.154%	70
몸 01ng	57	10	0.717%	44	91	0.140%	0.577%	81
하다 01ㄱvv	89	4	0.112%	46	86	0.146%	0.034%	82
너 01np	29	71	0.365%	707	2	2.250%	1.885%	69

전반적으로 각각 100 위 이내의 시어 중에서 모두 38 개의 시어가 공유되고 있다. 그러나 그렇다고 해서 두 시인의 시가 많이 다르다고는 할 수 없다. 그 이유는 두

가지이다. 하나는 앞에서도 살펴본 것처럼 김소월과 서지마의 공유 어종이 공유 어휘에 비해 그 비율이 매우 낮게 나타날 만큼 두 시인의 공유시어가 집중되어 있다는 점이다. 다른 하나는 중국어 언어표현, 특히 서지마의 언어 표현이 매우 다양하다는 것이다. 서지마는 동일한 하나의 의미를 여러 가지 다양한 시어를 사용해 표현하였다. 예를 들어, '길'에 관한 시어는 '道', '途径', '道儿' 등이 있다. 이와 같이 두 시인의 공유 시어 중 100 위 이내에 속하지 않으나 중요한 공유시어가 되는 경우도 있다. 따라서 두 시인의 시어 중에 100 위 이내에 속한 공유 시어를 결합하고 또, 중요한 시어를 뽑아 살펴볼 필요가 있다.

3. 결론

본고에서는 한•중 문학의 실증적 비교 연구, 특히 계량적 방법을 활용할 때 중국어 텍스트의 처리방법에 대해 논의해 보았다. 서지마 시 데이터 처리 사례를 들면서 김병선 교수님이 제시해 준 한•중•일 삼 개국 어휘 비교 방법론의 큰 틀에서 원전비평, 띄어쓰기, 번역문제를 중심으로 중국어 텍스트를 처리하는 구체적인 방법과 규칙을 정리하였다.

실증적 비교연구, 특히 계량적 어휘 비교 연구를 하려면 신빙성 있는 원전을 확정하는 것은 기본이다. 이로 인해 원문 판본 연구와 교열본 확정하는 작업을 우선해야 한다. 글자, 문장부호, 행 바꾸기 등 모두 원전비평에서 다루어야 하는 항목들이다.

중국어는 문장 단위로 되는 언어이다. 효율적인 어휘비교를 실현하기 위해 중국어문장을 어휘 단위로 띄어쓰고 어휘를 추출하는 작업을 해야 한다. 중국어를 띄어쓰기 할 때는 되도록 품사별로 띄어쓰되 실사, 허사를 나누고 합성어, 파생어, 속어, 외래어 등을 실제적인 상황을 보면서 띄어써야 한다.

한•중 어휘 비교 연구를 할 때 KoPoCo 데이터를 활용하려면 중국어 어휘를 한국어로 번역해서 처리하는 것은 효과적이다. 중국어를 한국어로 번역할 때는 품사별로 다의어, 비유어, 외래어 등을 구별하여 실제상황에 맞게 번역해야 한다.

한•중 양국 실증적 비교 연구는 시작하는 단계에 불과하다. 본고에서는 수월한 비교 성과를 내기 위해 중국어 텍스트를 한국어 텍스트와 같은 수준에 처리하는 방법을 구체적으로 논의했으나 이 방법론은 더 다듬어질 필요가 있다. 또한 서지마 말고 다른

중국 시인이나 작가의 텍스트를 처리하는 경험을 쌓으면서 중•한 양국 번역어 리스트를 작성하는 것이 필요하다. 한편으로는 본고에서 김소월과 서지마 시어 비교 수치를 여러모로 정리를 해 봤지만 이에 대한 분석이 아직 미흡한 상황이다. 추출된 시어의 동질성과 이질성에 대해 깊이 연구하는 것은 한•중 양국 현대문학 관련성과 발전 맥락을 밝히는 것에 일조가 될 것이다.

참고문헌

徐志摩文集 , 商務印書館, 1983.

徐志摩全集 , 廣西民族出版社, 1991.

徐志摩全集 , 五福出版社, 1976.

徐志摩全集 , 天津人民出版社, 2005.

김병선·조창환·배희숙·장노현 엮음, 『한국 현대시어 빈도 사전』, 한국문화사, 2007.

高偉, 「文學翻譯家徐志摩研究」, 上海外國語大學 博士論文. 2006.

具洸範, 「徐志摩的生平和思想簡論」, 華東師範大學 박사논문. 1996.

권소정, 「서지마 자연시 연구」, 고려대학교 석사논문. 2006.

권혜경, 「서지마 문학 연구」, 한국외국어대학교 박사논문 . 2002.

金相泰, 朴德垠, 文體의 理論과 韓國現代小說 . 한실, 1990.

김병선, 「어휘비교를 통한 한중일 근대문학 연구」, 제 5 회 세계한국학대회 발표문. 2010.

김병선, 국어와 컴퓨터 , 한실. 1992.

김병선, 소월의 시어와 그 쓰임새 , 한국문화사. 1994.

김병선, 한국 현대시어 빈도사전 , 한국문화사. 2007.

김종철, 「서지마시연구」, 성균관대학교 석사논문. 1985.

김해응, 「심연수 문학연구」, 한국학중앙연구원 박사논문. 2003.

만정정, 「김소월과 서지마의 시세계 비교연구」, 단국대학교 석사논문, 2009.

박안수, 「서지마 시 연구」, 영남대학교 박사논문. 2004.

박장례, 「박태원 소설의 문체 연구」, 한국학중앙연구원 박사논문. 2010.

朴燦惠, 「徐志摩 詩 研究」, 영남대대학교 석사논문, 1986.

왕해윤, 「韓國의 近代詩人 金素月과 中國의 近代詩人 徐志摩에 對한 比較 研究」, 경희대 학교 석사논문, 1974.

졸고, 〈김소월과 서지마 시 문학 비교연구〉, 한국학중앙연구원 박사논문. 2012.

유종호, 「옷과 밥과 자유」, 현대문학 . 2002.