어휘 비교를 통한 한중일 근대문학 연구 방안

Comparative Perspective to KCJ Modern Literature in terms of Literary Vocabulary

김병선(한국학중앙연구원)

1. 첫머리에

이 논문은 작품에 사용한 어휘를 중심으로 한중일 근대문학을 비교문학적으로 연구하는 방안에 대해 연구한다. 그 동안 비교문학 연구는 크게 보아 유럽식의 실증주의적 영향관계의 탐구나, 미국식의 일반문학적인 유사현상의 탐구 등을 주류로 하여 전개되어 왔다. 이 연구는 시기를 근대문학시기로 한정하고, 한중일 삼국의 대표적 근대시인의 작품 세계를 비교 연구하되, 그 어휘에서 나타나는 계량적 현상에 주목하여 상호 비교하는 여러 가지 방법을 논의해 보고자 한다.

고전문학 시기의 한중일 삼국 문학의 직접적 영향관계에 대해서는 그 동안 많은 연구가 이루어졌다. 이 세 나라는 지리적으로 매우 가깝고, 문학을 포함한 문화 전반의 교류가 활발하게 이루어져 왔던 것이 사실이다. 근대에 있어서는, 근대화에 한 발 늦은 한국의 문학은 일본이나 중국을 거쳐온 서구문학의 영향을 받았다. 많은 한국의 문인들이 이 시기에 일본이나 중국 유학을 했으며, 공적인 학교 교육을 통해서 혹은 개인적인 사숙(私塾) 혹은 독서활동 등을 통해서, 서구문학 및 서구문학의 영향을 받은 현지문학의 영향을 입었다.

이와 같은 이유로 영향관계의 탐구를 통한 한중일 근대문학에 대해서는 현재 여러 가지 연구가 진행되고 있으나, 정작 작품 자체의 연구, 작품을 통해 주고 받은 언어에 대한 연구는 매우 미흡한 상황이다. 우리나라 근대 어휘의 형성과정에 일본과 중국과의 문명교류가 차지하는 비중이 높고, 문학의 경우에도 번역과 번안의 영향이 컸다고 할 수 있다. 따라서 영향관계를 주고받았던 작품들, 작가들 사이의 어휘적 유사성을 발견할 수 있으리라 생각한다. 이 연구에서는 어휘의 비교에 주목하기 때문에 시 장르에 한정하여 논의하려고 한다. 시는 운문으로 사상과 정서를 압축적으로 표현하며, 어휘 하나하나 및 토씨 하나하나에도 심혈을 기울이기 때문에, 시 연구에 있어서는 어휘나 어법에 대한 연구와 더불어 어휘의 계량적 현상에 대한 탐구도 매우 중요하다.

작품의 어휘에서 나타나는 계량적 현상에 대한 연구는 그 동안 본격적으로 이루어진 바 없다. 사실상 특히 한국에서 계량적 현상에 대한 연구 자체가 이제 시작되는 상황이며, 계량 연구를 위한 텍스트의 확보도 이제야 어느 정도 이루어지고 있는 상황이다. 아울러서 계량적 현상에 대한 분석의 기법과 이를 문학적 해석으로 확장하는 것도 일부 학자들만 시범적으로 하고 있다. 아울러서 하나의 언어가 아닌 두 개 이상의 언어와 문학을 비교한다는 것은 개인 연구자에게는 큰 부담이 아닐 수 없다. 두 개의 문학적 언어에 대한 지식이 충분해야 하고, 상호 비교에서 파생되는 각종 문제에 대한 대응력이 필요하기 때문이다. 이 연구는 이러한 여러 가지 난점에도 불구하고, 한중일 삼국 문학 언어의 비교를 시도해 볼 것이며, 주로 현재 수행하고 있는 한-일 어휘 비교의 구체적 현황을 소개하고, 한-중에 대해서는 그 가능성을 타진하기로 한다.

2. 문학 연구를 위한 코퍼스의 구성

필자가 한중일 삼국 문학 어휘에 대한 연구를 시작하게 된 것은, 한국 근대문학 코퍼스를 구축하여 어휘 자료가 축적되어 있고, 이를 바탕으로 몇 가지 계량적 연구를 수행하면서 이를 비교문학 연구에 시도해 보고자 하는 의욕에서다. 아울러서 필자가 학위과정

에서 지도하고 있는 외국인 학생들에게 비교문학적 주제의 학위논문 지도를 하면서, 이 학생들의 언어적 능력을 활용하고, 필자의 코퍼스와 분석 기법을 적용하면 매우 좋은 결과가 나올 것으로 생각했기 때문이다.
현재 한국 근대문학 어휘에 대한 자료 축적과 계량적 연구에 관해서는 필자가 소속하고 있는 한국학중앙연구원이 선도적 입장에 있다. 그 동안 현대시 데이터베이스 구축사업을 10 년 가까이 추진하여 『한국현대시어 용례사전』과 『한국현대시어 빈도사전』 등을 출판한 바 있으며, 최근에는 신소설 코퍼스 확보를 통해, 신소설 어휘사전 편찬 사업을 수행하고 있다. 이 신소설 어휘사전 편찬에는, 주로 1900 년대부터 1910 년대 말까지 단행본으로 출판된 창작 신소설을 우선하고, 번안·번역 작품 중에서도 중요 작품을 연구 대상으로 포함하였다. 3 차년도까지 60 편 규모까지 확장하여 약 1 백만 어절 규모의 코퍼스를 구축할 것이다.

가. KoPoCo 의 범위[1]

KoPoCo 에는 1923 년부터 1950 년까지 발간된 창작 시집에 수록된 현대 창작시가 수록되어 있다. 사실 KoPoCo 는 주요 항목이 마크업된 코퍼스 형식으로 되어 있기도 하지만, 실제로는 용례사전을 중심으로 하는 각종 테이블이 MS Access 데이터베이스 형식으로 운영되고 있다. 한국 현대시 코퍼스에는 원전 비평을 거쳐 교정본으로 확정된 현대시 8,201 편을 수록하였으며, 수록 대상이 되는 시집은 모두 196 종, 시인 수는 모두 345 명에 이른다.

나. KoPoCo 의 구성

- 어휘 기본형: 각 활용형(곡용형) 어절에 대하여 형태소 분석을 통해 기본형(원형)을 제시한다. 이 기본형 어휘를 대상으로 동음이의어 및 다의어 분석을 하고, 국립국어원의 『표준국어대사전』에 근거하여 첨자로써 표시한다. 이러한 분석 과정에서 품사도 함께 분석한다.
- 어휘 구성 분석: 기본형 어휘의 성격과 구성 방식을 기호로 나타낸다. 한국어(고유어, 한자어, 기타), 외래어, 외국어 등을 구별하여 각 어휘마다 그 결합 방식을 표시하였다.
- 발표 연도: 원전비평의 과정을 거치면서 1923 년부터 1950 년 사이에 두 번 이상 발표된 작품 중 신빙성이 높은 한 편만을 KoPoCo 에 수록하였다. 대체로 시인의 생전에 발표된 경우는 나중에 발표된 것을 우선으로 하였고, 공동시집보다는 단독시집을 우선으로 하였다.

3. 비교 연구 방법론

가. 비교 대상 작가와 작품 선정

---

[1] KoPoCo 는 한국 현대시 코퍼스(Korean Modern Poetry Corpus)에서 일부 단어의 첫 두 글자를 따서 만든 말이며, 뒤의 숫자는 해당 연도의 코퍼스를 뜻한다. KoPoCo 에 관해서는 필자(2004)가 「한국 현대시 데이터베이스의 구성과 그 활용 방안」(『한국언어문학』 제 53 집. pp.513-535.)에서 그 대체적인 구성 내용을 소개했고, 또 Kim, Byongsun(2005), The Present Conditions and Tasks in Constructing the Database of Korean Literary Materials Centering on the Korean Poetry Corpus (*The Review of Korean Studies*, 2005 Winter*8-4). pp.105-139.), 김병선(2006), 「현대시인의 문체적 지문을 찾아서」, 『국어국문학』 제 143 호, 국어국문학회. pp.153- 188. 등에서 그 구축 방법에 대해 자세히 논의한 바 있다. 그 이후 표제어 확정에 변화가 있었고, 어휘에 대한 추가적 분석이 이루어졌다. 여기서는 최근의 현황을 소개한다.

한중일 삼국 대표시인의 작품에 사용된 어휘 비교를 위해서는 먼저 작가와 작품을 선정해야 한다. 비교문학적으로 영향관계가 있는 시인을 상호 비교하는 것도 의미가 있겠지만, 이 연구는 이러한 분야의 첫 시도이기 때문에 한중일 삼국의 대표시인을 상호 비교하기로 한다. 한국에서는 김소월(1902-1934)을 선정하기로 한다. [2]
일본의 작가로는 이시카와 다쿠보쿠(石川啄木, 1886~1912), 중국의 서지마(徐志摩, 1896~1931)를 대상 작가로 선정하였다.[3] 두 작가 모두 한국 문단과의 관계는 있지만, 시작품에서 실제적 영향관계가 있는 한국 시인은 발견하기 어렵다. 따라서 근대문학기의 각 나라의 대표적 시인이라는 평판에 근거하여 세 시인을 선정한 것이다. 다쿠보쿠는 일본의 대표적인 단가 문학의 시인이라는 평을, 서지마 역시 중국 근대시인 중의 최고봉이라는 평을 받는다.
비교 연구에 있어서는 일반적으로 비교 대상 각각의 크기가 클수록 비교의 타당성이 높아진다. 다쿠보쿠의 작품은 비교적 원문 교정이 잘 되어 있는 『石川啄木全集』(筑摩書房, 1978) 판본으로 택하였고, 이 중에서 다쿠보쿠의 시세계를 제일 잘 나타내고 있는 대표적 시집인「동경」,「한줌의 모래」,「호루라기와 휘파람」, 및「슬픈 장난감」등 네 개 시집에 실려 있는 시와 단가만을 연구대상으로 한정하였다. 서지마의 경우는 몇 개의 전집들이 나와 있는데, 판본 비교를 통해서 가장 정확도가 높은 '천진판(天津版)'을 기본으로 하고, 일부 오류는 다른 판본에 근거하여 바로잡았다.

나. 비교를 위한 어휘의 추출

1) 대상 작품의 전산 입력

소월의 작품은 KoPoCo 에 이미 각 어휘가 추출되어 있고, 계량적 처리도 해놓은 상태이므로, 다쿠보쿠와 서지마 텍스트를 위와 같은 수준에 이르게 하는 것이 기본적 전산 입력의 목표다. 우선 선정한 텍스트를 일단 워드프로세서 프로그램으로 입력하였다. 소월을 비롯한 한국문학 작품의 전산입력에는 <아래한글>(한글과컴퓨터사)을 사용하였지만, 서지마와 다쿠보쿠의 텍스트는 각각 일본어와 중국어로 되어 있으므로, 해당 외국어를 입력하기 수월한 워드프로세서를 사용하도록 하였다. 그래서 MS Word 의 중문판과 일문판을 사용하였다. <아래한글>과 MS Word 의 파일 형식이 다르기는 하지만, 모두 유니코드 기반의 워드프로세서이기 때문에 최종적으로는 UTF-8 텍스트 파일로 저장하여 코드체계를 맞추어 주었다.

2) 텍스트 입력 후처리

시작품의 어휘를 추출하는 데 있어서 한국어 전산 파일은 매우 유리한 상황에 있다. 말하자면 한국어 텍스트는 어절별로 띄어쓰기를 하기 때문에 각 어절에서 기본형을 추출하면 되는 것이다. 그에 비해 중국어 텍스트와 일본어 텍스트는 문장 단위의 띄어쓰기만 하고 있어서 이로부터 각 어휘를 추출하기 위해서는 후처리(post processing)를 거쳐야 한다.

---

[2] 중국과 일본의 작품에 대해서는 필자가 학위과정에서 지도하고 있는 학생들과 공동작업을 하기로 하였다. 다쿠보쿠의 작품은 일본 국적의 하야시 요코(林陽子)가, 서지마의 작품은 중국 국적의 요위위(姚委委)가 입력과 교정 및 원전확정을 담당하였다.
[3] 이 연구에서는 편의상 김소월은 '소월'로, 이시카와 다쿠보쿠는 그의 필명인 '다쿠보쿠'로 표기하기로 한다. 자료처리를 위한 약어로 소월은 'S'로, 다쿠보쿠는 'T'로 표시한다. 서지마는 한국 한자음으로 표기하되, 약어로는 'J'으로 표시한다.

또한 어휘 단위로 띄어쓰기를 하는 데 있어서도 그리 간단하지는 않았다. 비교 대상이 한국어 어휘들이기 때문에 한국어 어휘와 일대일로 대응될 수 있도록 띄어쓰기를 해야 하나, 각 언어에 고유한 연어 등이 이러한 작업에 애로를 가져왔다. 이런 경우는 최선을 다해 어휘 별로 분석하였지만, 그렇게 할 수 없는 어휘들은 그대로 두었다.
어휘 분리를 한 다음에는 일차적으로 용례색인(concordance)을 만들어야 하는데, 이 색인에서 충분한 문맥 정보를 보이기 위해서, 시 본문의 행 구분은 빗금(/)으로, 연 구분은 쌍빗금(//)으로 치환해 두었다.

3) 용례색인의 생성

입력된 텍스트는 해당 워드프로세서에서 텍스트 파일(*.txt)로 저장한 다음, 필자가 제작한 <똑똑새> 프로그램을 통해서 용례색인을 자동으로 생성하였다. 용례색인에는 ID(원문의 출현 순서 번호), 기본형(최초에는 활용형 어휘가 제시되며, 별도의 과정을 거쳐 기본형 표제어를 제시한다.), 활용형(원문에 쓰인 어절 키워드),[4] 앞 문맥(키워드의 앞쪽에 있는 문맥), 뒤 문맥(키워드의 뒤쪽에 있는 문맥), 제목(작품의 제목), 출전(작품의 출전 시집과 간행 연도), 기타(헤더 부분에 마크업 된 정보) 등의 정보가 수록된다.

4) 어휘별 번역 처리

한중일 삼국은 언어가 서로 다르고, 사용하는 문자체계도 다르다. 같은 한자문화권이어서 삼국이 공유하고 있는 어휘가 적지는 않지만, 더러는 뜻은 같더라도 표기가 다른 어휘라든지, 표기는 같지만 뜻이 다른 어휘도 있어서 오히려 번역에 장애가 되기도 한다.(예: 東西: 한국-동쪽과 서쪽, 중국-물건)
일단 언어와 문자가 다른 언어권의 문학작품에 사용된 어휘를 비교하려면 기준이 되는 언어로 번역해야 한다. 즉 비교 대상이 되는 출발 문학작품을 하나의 언어로 번역하고, 목표 문학작품도 같은 언어로 번역한 다음에 서로 비교하는 것이다. 여기서 말하는 '하나의 언어'는 일종의 중간언어로서, 자연어일 수도 있고 인공어일 수도 있다. 일반적으로 어휘 비교에서 이러한 중간언어는 비교문학자의 모국어가 될 가능성이 높다. 필자도 중국 작품을 한국어로 번역하고, 일본 작품도 한국어로 번역하여 상호 비교하고 있다. 또 이러한 비교방법은 당연히 한국 작품과 비교하는 경우에 효과적이다.
필자의 경우는 한국어를 모국어로 하는 사람으로서 개인적으로 중국어와 일본어에 대해 약간의 이해력 정도만 가지고 있다. 따라서 필자는 비교연구를 주제로 학위논문을 작성하고 있는 외국인 학생들과 공동 작업으로 번역을 하기로 하였다. 이 외국인 학생들은 자기 모국어와 한국어를 대단히 우수하게 구사하는 사람들인데, 이러한 여러 가지 면에서 중간언어로서 한국어를 택하는 것이 유리한 점이 있었다.
번역에 있어서는 작품 전체의 번역 방법과 어휘의 번역 방법이 있다. 작품 전체를 번역하는 것이 문맥상의 의미를 제대로 살릴 수 있는 번역 방법이어서 가장 바람직하지만, 워낙 품이 많이 드는 것이 문제다. 작품 전체를 번역하면서 문맥적 의미를 살리다 보면 의역 쪽으로 갈 가능성이 높은데, 실제 어휘 비교에서는 작품 내부적인 의미보다는 표현 자체의 사전적·기본적 의미 쪽에서 처리의 수준을 정하는 것이 현실적이라고 생각한다. 따라서 일단 용례색인을 만들어 키워드 기준으로 정렬한 이후에, 해당 키워드를 번역하는 방법을 택했고, 이때 의미가 모호한 경우에는 문맥 정보를 참조하는 수준으로 처리하

[4] 기본형과 활용형(혹은 곡용형)은 주로 한국어와 일본어 같은 교착어(뜻을 나타내는 말에 문법적 관계를 표시하는 말이 덧붙는 언어)나 일부 굴절어에서 구분되는 것이다. 중국어 같은 고립어(문장 속에서 단어의 위치에 따라 그 단어가 문법적 구실을 하는 언어)는 일부 어조사를 제외하면 특별히 활용형을 말하기는 어렵다.

였다. 즉 어휘에 대한 개별적 번역의 방법을 택하여 객관성과 효율성을 높였다고 할 수 있다.

다. 어휘 번역의 문제와 그 처리

1) 한국어 의미어의 부여

어휘 비교에 있어서는 표준적인 어휘를 설정하고, 그것을 기준으로 삼아 비교해야 한다. 한 사람의 시인이 작성한 텍스트 내부에도 적지 않은 수의 이음동의어(synonym) 혹은 유의어가 나타난다. 그런데 이것을 각각 별개의 어휘로 처리한다면 특히 비교문학적 연구에서는 매우 많은 문제를 야기할 수 있다. 따라서 이음동의어를 표준화하여 처리할 수 있는 방안을 마련해야 한다.
KoPoCo 어휘의 기본형을 설정하는 데도 이것이 문제가 되었다. 시인들은 기본적으로 다양한 어휘를 사용하는 것을 기본적 업무로 하는 사람들이어서, 표준어를 고집하기 보다는 지역어(사투리, 방언)나 옛말 등을 살리려 애를 쓰고 있으며, 심지어는 새로운 어휘를 조합해 내기도 한다. 이러한 비표준적인 어법들이 시어의 독특성을 나타내는 부분이기 때문에 이를 무시할 수가 없다. 따라서 필자는 시어의 독특성과 일반성을 적절하게 연결할 수 있는 방안으로서 '의미어(semantic term)'라는 것을 각 시어에 적용하였다. 어절에 대한 형태소 분석을 통해 기본형을 추출할 때에 가급적 시인의 표기형을 살리되, 이 의미어 필드에는 각 표기형에 대한 표준적인 관련어(associate term)를 제시한 것이다. 예를 들면, '태양'과 '해', '햇님', '해님'에 대하여 '태양'을 의미어로 하는 것을 말한다. 이를 통해서 유의어들이 하나의 표준적인 어휘로 대치된다. KoPoCo 의 모든 어휘에 대해서 이 의미어가 제시되어 있다.

2) 비유어와 다의어의 처리

만약 비유적 표현이라면 어떻게 할 것인가? 의역 경우에는 이런 경우에는 비유의 원관념(tenor)으로 번역하겠지만, 여기서는 객관성을 살리기 위해서 보조관념(vehicle)을 그대로 제시하였다. 예를 들어 영어 단어 'sun'은 중국어로 '太陽', '(위성을 가진) 恒星' 등의 기본적 의미 외에 비유적 의미로는 '중심인물, 영예, 권력'으로, 시적으로는 '하루 혹은 일년' 혹은 '(구식의) 버너' 등을 가질 수 있다. 그러나 이 비교연구에서는 오직 '태양'이라는 어휘로만 표준화하여 번역하기로 한다.
아울러서 이런 말은 다의어(polysemy)의 성격을 가지고 있기도 한데, 다의어의 경우는 맥락을 고려하여 적절한 번역어를 제시하기로 한다. 위의 예에서는 어떤 경우에는 '태양'으로, 다른 문맥에서는 '항성'으로 번역하는 것이다.

4) 기본형 설정의 범주

한국현대시 코퍼스에서는 어절 단위로 용례색인을 만들었기 때문에 조사나 어미의 경우는 별도의 키워드로 제시하지 않았다. 그런데 일부 일본 문장에서는 이를 구분할 필요도 있었다. 따라서 일본어 텍스트에서 일부 조사와 어미를 목록에 올리기는 하지만 비교 대상에서는 제외하기로 한다.
중국어 텍스트에서 조사(구조조사 '的'.'地'.'得'.'似的'.'所' 등, 시태조사 '了'.'着'.'過' 등, 어기조사 '嗎'.'呢'.'吧'.'啊' 등)는 붙여쓰기로 하였다. 따라서 '辞别了','逃出了'등으로 처리하여, 한국어의 어휘와 대등하도록 하였다. '要 香烟吗''是 故乡吗'도 같은 식으로 처리했다. 방위와 위치를 가리키는 방위사(上, 下, 前, 后, 左, 右, 东, 西, 南, 北, 里,

外, 中, 內, 旁 등)의 경우는 앞의 어휘와 띄어쓰기로 하였다. '大殿 里''靜定 中'도 한국어 대응성을 고려하여 띄어쓴 것이다.
복수형의 경우는 접사로 보아서 이를 단수 표준형으로 제시하였다. 그러나 중국어의 '我們(우리)'처럼 한국어에 복수형의 별도 어휘가 대응되는 경우에는 하나의 단어로 보아 그대로 두었으나, 나머지는 단수형으로 제시하였다.

4. 한-일 시어 비교 연구

가. 소월, 다쿠보쿠의 시형식의 일반적 특성

한-일 시어를 비교하기 위해서 소월과 다쿠보쿠의 시작품 텍스트를 전산 처리하여 몇 가지 계량적 분석을 시도해 보았다. 먼저 시형식에 대한 일반적 특성을 분석하기로 한다.

표 1 소월, 다쿠보쿠, KoPoCo 일반통계

	작품수	어종	어휘	어휘/어종 반복지수	작품당 어휘수
소월	145	2,152	7,949	3.694	54.82
다쿠보쿠	832	3,964	18,936	4.777	22.76
KoPoCo	8,201	34,882	611,884	17.542	74.61

작품당 어휘수 면에서 소월은 다쿠보쿠보다 2 배 이상으로 나온다. 소월 역시 한국현대시에서는 비교적 짧은 작품을 쓴 시인인데, 이러한 현상이 나오는 것은 다쿠보쿠의 텍스트가 상당 부분 단형시가 양식으로 되어 있기 때문이다.
어휘/어종의 반복지수를 분석한 결과 소월보다 다쿠보쿠의 반복지수가 상당히 높은 것으로 나온다. 반복지수란 한 종류의 시어를 평균적으로 몇 차례나 사용했는가를 평가하는 지수로서, 일반적으로 표본이 많을수록 반복지수는 다소 높아진다. 평가 결과 다쿠보쿠는 어휘 다양성 면에서 소월보다 낮은 것으로 판단된다.
소월과 다쿠보쿠 그리고 한국현대시의 품사 분포에 대해서도 분석해 보았다.

표 2 KoPoCo, 소월, 다쿠보쿠 품사 빈도

품사 필드	KoPoCo 빈도	K 비율	S 빈도	S 비율	T 빈도	T 비율
명사	250,928	41.009%	3,184	40.055%	9,107	48.094%
대명사	32,928	5.381%	466	5.862%	828	4.373%
의존명사	16,909	2.763%	177	2.227%	126	0.665%
고유명사	7,106	1.161%	74	0.931%	99	0.523%
수사	1,748	0.286%	15	0.189%	29	0.153%
동사	149,215	24.386%	2,120	26.670%	4,823	25.470%
보조동사	15,534	2.539%	150	1.887%	324	1.711%
형용사	58,802	9.610%	736	9.259%	1,643	8.677%
보조형용사	2,253	0.368%	28	0.352%	118	0.623%
부사	45,454	7.429%	649	8.165%	931	4.917%
관형사	23,672	3.869%	278	3.497%	354	1.869%

접미사	88	0.014%	1	0.013%	5	0.026%
접두사	77	0.013%	0	0.000%	16	0.084%
감탄사	4,425	0.723%	57	0.717%	182	0.961%
미상	482	0.079%	2	0.025%	32	0.169%
어근	424	0.069%	3	0.038%	1	0.005%
어미	3	0.000%	0	0.000%	76	0.401%
조사	1,836	0.300%	9	0.113%	241	1.273%
	611,884	100.000%	7,949	100.000%	18,936	100.000%

소월의 경우는 한국현대시 모집단과 비교해 볼 때, 각 품사별로 거의 비슷한 분포를 보이고 있다. 동사의 비율이 약간 높기는 하지만 크게 두드러지는 것은 아니다. 보조동사와 함께 계산한다면 거의 같게 될 것이다. 각 품사별로 현상을 분석하면 다음과 같다.
- T 명사 빈도가 상당히 높다. 품사 확정 과정에서 명사로 처리했던 것을, 명사형 어미('ㅁ')가 붙은 것을 명사가 아니라 형용사나 동사로 바꾸었는데도 높다. 이러한 현상이 언어의 특수성에서 기인하는지, 시의 표현 기법에서 비롯된 현상인지는 차후 더 살펴볼 여지가 있다고 본다. 사실 작품 양을 고려한다고 하더라도 사용하고 있는 명사의 종류도 다쿠보쿠가 훨씬 많다. (소월 945 종 3,184 개(명사 반복지수 3.369), 다쿠보쿠 2,349 종 9,107 개(명사 반복지수 3.877))
- T 고유명사 빈도가 상대적으로 낮다.
- T 동사 빈도는 S 빈도와 거의 비슷하다.
- T 형용사 빈도는 약간 낮다.
- T 부사 빈도는 상당히 낮다.
- 기타 조사, 어미, 접미사, 접두사 등 교착어에서 기능을 당하는 부분에 대해서는 특별히 의미를 두기 힘들다고 본다. 그리고 빈도가 높지도 않을 뿐 아니라, 두 시인 사이의 차이도 그리 크지 않다.

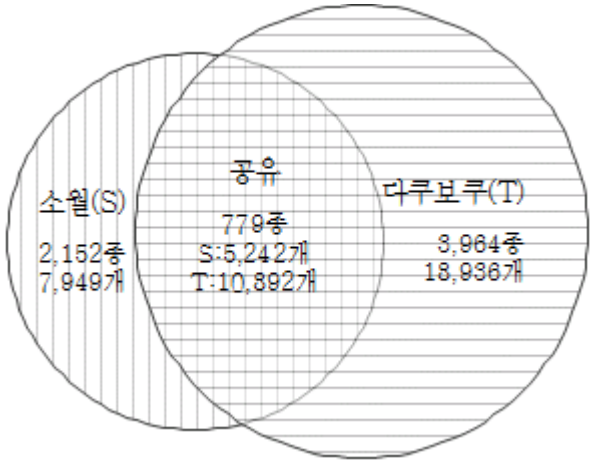
나. 소월-다쿠보쿠 유사성 검토

소월과 다쿠보쿠 시어의 유사성을 검토하기 위해서 먼저 공유시어(shared word)를 추출하였다. 공유시어란 비교 대상이 되는 두 시인의 작품에 모두 나타나는 어휘를 말한다. 공유시어를 시어의 출현빈도와 더불어 검토하면 시인 사이의 유사성을 밝히는 하나의 방법이 될 수 있다.

공유시어를 추출하기 위해서 다음과 같은 절차를 밟았다.
- 각 시어 활용형의 기본형을 추출하였다.
- 각 시어에 대해 품사 분석을 하였다.
- 각 시인의 문맥에서 중의성을 가진 시어는 동음이의어 분석을 하였다.
- 적절한 컴퓨터 처리를 통해서 두 시인의 전체 시어(어종) 목록을 만들고, 시인별로 출현 빈도 필드에 쿼리를 통해서 각각의 빈도수를 추가하였다.

그 결과 S 빈도(소월 시의 빈도)와 T 빈도(다쿠보쿠 시의 빈도)에 공통으로 1 이상의 값이 들어 있는 어휘를 공유시어로 설정하였다.

표 3 소월-다쿠보쿠 시어 공유 도표



계산의 결과를 도표로 나타내면 다음과 같다.

표 4 어종-어휘별 공유, 비공유 통계

	어종			어휘		
	공유	비공유	계	공유	비공유	계
전체	779 종	4,558 종	5,337 종	16,134 개	10,751 개	26,885 개
	14.60%	85.40%	100%	60.01%	39.99%	100%
소월	779 종	1,373 종	2,152 종	5,242 개	2,707 개	7,949 개
	36.20%	63.80%	100%	65.95%	34.05%	100%
다쿠보쿠	779 종	3,185 종	3,964 종	10,892 개	8,044 개	18,936 개
	19.65%	80.35%	100%	57.52%	42.48%	100%

* 공유 어휘의 비율이 높고, 비교 대상 시인의 비율이 비슷할 때, 전반적으로 시인의 유사성은 높아진다고 평가할 수 있다.
* 공유와 비공유의 면에서 어종과 어휘의 비율을 살펴보니, 전반적으로 공유 어종이 공유 어휘에 비해 그 비율이 매우 낮게 나타났다. 소월의 경우는 공유 어휘는 65% 이상인데 비해, 공유 어종은 36% 정도이고, 다쿠보쿠는 그 편차가 더 커서 공유 어휘가 57% 이상인데 비해, 공유 어종은 20%에 미치지 못한다. 이러한 현상은 각 시인이 집중적으로 사용하는 어휘들의 공유성이 높다는 것을 말한다.

이제 공유시어를 구체적으로 살펴보기로 한다.
이 연구에서는 중요 공유시어(significant shared word)를 선정하기 위해서 다음과 같은 두 가지 실험을 진행하였다. 하나는 해당 시어가 전체 시어에서 차지하는 비율을 측정하여 양 시인이 일정 수준 이상의 비율로 사용한 시어만을 선정하는 것이다. 다음 표에서 S 비율은 S 빈도를 소월의 전체 시어(7,949 개)로 나눈 백분율값이고, T 비율은 다쿠보쿠의 경우다. 그리고 그 백분율값 차이의 절대값을 비율차로 계산하였다. 이렇게 할 경우에 빈도가 아주 낮은 어휘들의 비율차가 매우 적은 것으로 나타나 원하는 결과를 얻을 수 없었다.
따라서 또 다른 방법으로 순위를 따지는 방법을 적용하였다. S 순위는 소월 전체 시어 중 S 빈도값으로 순위를 매긴 것이고, T 순위는 다쿠보쿠의 경우다. 그리고 그 순위의 차이를 절대값으로 표시한 것이 순위차 필드다. 다음 표는 그 순위의 차이가 공히 100 위 이내인 65 개의 시어를 선정하고, 그 차가 적은 것부터 오름차순으로 정렬한 것이다.

표 5 중요 공유시어 (S순위, T순위 모두 100위 이내 중 순위차 20인 것)

시어	S 빈도	S 순위	S 비율	T 빈도	T 순위	T 비율	비율차	순위차
나 03np	187	1	2.352%	395	1	2.087%	0.266%	0
얼굴 01ng	14	92	0.176%	30	93	0.158%	0.018%	1
사람 00ng	52	12	0.654%	116	9	0.613%	0.041%	3
세상 01ng	34	31	0.428%	67	34	0.354%	0.074%	3
깊다 00vx	15	84	0.189%	32	80	0.169%	0.020%	4
날 01ng	44	15	0.554%	103	19	0.544%	0.009%	4
고향 02ng	24	47	0.302%	43	51	0.227%	0.075%	4
하늘 01ng	39	22	0.491%	78	26	0.412%	0.079%	4

그 01mm	61	9	0.767%	108	15	0.571%	0.197%	6
밤 01ng	54	11	0.679%	107	17	0.565%	0.114%	6
알다 00vv	25	44	0.315%	61	38	0.322%	0.008%	6
불다 01vv	19	61	0.239%	42	55	0.222%	0.017%	6
부르다 01vv	20	56	0.252%	38	64	0.201%	0.051%	8
저 04mm	38	24	0.478%	69	32	0.365%	0.114%	8
이 05mm	41	18	0.516%	116	9	0.613%	0.097%	9
오다 01vv	104	2	1.308%	114	12	0.602%	0.706%	10
서다 01vv	19	61	0.239%	43	51	0.227%	0.012%	10
땅 01ng	15	84	0.189%	33	74	0.174%	0.014%	10
생각 01ng	16	78	0.201%	31	88	0.164%	0.038%	10
꽃 01ng	27	39	0.340%	75	29	0.396%	0.057%	10
없다 01vx	44	15	0.554%	166	2	0.877%	0.323%	13
소리 01ng	40	20	0.503%	129	7	0.682%	0.178%	13
길 01ng	28	36	0.352%	44	49	0.232%	0.120%	13
있다 01ⓔvx	17	74	0.214%	31	88	0.164%	0.050%	14
눈물 01ng	22	50	0.277%	38	64	0.201%	0.076%	14
달 05ng	24	47	0.302%	39	61	0.206%	0.096%	14
당신 02np	28	36	0.352%	86	22	0.454%	0.102%	14
꿈 01ng	40	20	0.503%	138	5	0.729%	0.226%	15
무엇 00np	18	70	0.226%	42	55	0.222%	0.005%	15
내리다 01vv	17	74	0.214%	41	58	0.217%	0.003%	16
너 01np	20	56	0.252%	57	40	0.301%	0.050%	16
하다 01vv	89	4	1.120%	90	21	0.475%	0.644%	17
아침 00ng	14	92	0.176%	33	74	0.174%	0.002%	18
그 01np	26	43	0.327%	39	61	0.206%	0.121%	18
마음 01ng	39	22	0.491%	156	3	0.824%	0.334%	19
들 01ng	19	61	0.239%	32	80	0.169%	0.070%	19
가다 01vv	99	3	1.245%	83	23	0.439%	0.807%	20
노래 01ng	17	74	0.214%	43	51	0.227%	0.013%	23
죽다 01vv	38	24	0.478%	45	47	0.238%	0.240%	23
보다 01vv	35	30	0.440%	134	6	0.708%	0.268%	24
그림자 00ng	14	92	0.176%	37	68	0.195%	0.019%	24
바다 00ng	25	44	0.315%	36	70	0.190%	0.124%	26
때 01ng	78	5	0.981%	70	31	0.370%	0.611%	26
오늘 00ng	19	61	0.239%	31	88	0.164%	0.075%	27
또 00ma	21	52	0.264%	83	23	0.439%	0.174%	29
울다 01vv	71	6	0.893%	64	37	0.338%	0.555%	31
혼자 01ng	14	92	0.176%	40	60	0.211%	0.035%	32
속 01ng	19	61	0.239%	30	93	0.158%	0.081%	32
사랑 01ng	19	61	0.239%	76	28	0.402%	0.162%	33
흐르다 01vv	27	39	0.340%	33	74	0.174%	0.165%	35

가다 01vx	19	61	0.239%	29	97	0.153%	0.086%	36
되다 01vv	49	13	0.616%	44	49	0.232%	0.384%	36
몸 01ng	57	10	0.717%	46	46	0.243%	0.474%	36
집 01ng	19	61	0.239%	28	98	0.148%	0.091%	37
일 01ng	18	70	0.226%	68	33	0.359%	0.133%	37
바람 01ng	36	27	0.453%	38	64	0.201%	0.252%	37
봄 01ng	41	18	0.516%	42	55	0.222%	0.294%	37
가슴 01ng	20	56	0.252%	106	18	0.560%	0.308%	38
것 01nb	43	17	0.541%	41	58	0.217%	0.324%	41
듣다 01vv	14	92	0.176%	50	44	0.264%	0.088%	48
눈 01ng	15	84	0.189%	73	30	0.386%	0.197%	54
물 01ng	29	34	0.365%	28	98	0.148%	0.217%	64
아아 01ic	16	78	0.201%	128	8	0.676%	0.475%	70
잊다 01vv	36	27	0.453%	28	98	0.148%	0.305%	71
살다 01vv	63	8	0.793%	30	93	0.158%	0.634%	85

전반적으로 각각 100 위 이내의 시어 중에서 모두 65 개의 시어가 공유되고 있다는 것은 두 시인의 어휘가 매우 유사하다는 평가를 할 수 있게 만든다.
두 시인 모두 서정시인으로서 각 제 1 위 시어인 '나[我]'가 중요 공유시어에서 최상위를 차지하는 것은 당연한 일이라 할 수 있다. 서정시는 시적 화자의 정서와 생각을 읊는 것이기 때문이다. 한국어의 문법적 영향 하에 있는 한국시에서는 '나'라는 주어가 생략될 수 있기 때문에 김소월의 '나'는 시적 화자, 시적 주체의 적극적 노출이라고 분석될 수도 있다.
중요 공유시어들을 품사별로 살펴보기로 한다.
- 대명사 시어로는 1 인칭, 2 인칭, 3 인칭 등의 어휘가 중요 공유시어로 올라 있다.(나 03np, 당신 02np, 너 01np, 그 01np, 무엇 00np) 특기할 만한 것으로는 미확정 대상을 가리키는 지시대명사인 '무엇'이 여기에 포함되어 있다는 것이다. 세계와 사물에 대한 궁금함이 시적 모티브가 되고 있지 않는가 추정할 수 있다.
- 명사의 빈도가 전체적으로 높은 만큼 중요 공유시어도 명사 어휘가 많다. 대체적으로 인간사와 관련된 어휘들이 주축을 이루고 있다. (얼굴 01ng, 사람 00ng, 눈물 01ng, 소리 01ng, 몸 01ng, 가슴 01ng, 눈 01ng, 생각 01ng, 꿈 01ng, 마음 01ng, 노래 01ng, 사랑 01ng, 혼자 01ng, 속 01ng, 일 01ng) 삶의 기본 조건이 되는 자연 환경(세상 01ng, 하늘 01ng, 땅 01ng, 달 05ng, 바다 00ng, 바람 01ng, 꽃 01ng, 그림자 00ng, 물 01ng) 및 시기에 관한 어휘(날 01ng, 밤 01ng, 아침 00ng, 봄 01ng, 때 01ng, 오늘 00ng)가 적지 않다. 이외에 지리적 배경과 관련되는 어휘(길 01ng, 들 01ng, 고향 02ng, 집 01ng)들도 보인다. 명사의 경우 대체적으로 일반적인 빈도가 높은 어휘라고 볼 수 있으며, 특별히 문학어로서 거론할 만한 어휘는 보이지 않는다.
- 동사 어휘로는 인간의 기본적 활동과 관련된 어휘(알다 00vv, 부르다 01vv, 오다 01vv, 서다 01vv, 있다 01ⓒvx, 내리다 01vv, 하다 01vv, 가다 01vv, 가다 01vx, 죽다 01vv, 보다 01vv, 울다 01vv, 듣다 01vv, 잊다 01vv, 살다 01vv)가 많은 비중을 차지하고 있고, 자연물의 움직임과 관련된 어휘(불다 01vv, 흐르다 01vv, 되다 01vv)도 일부 보인다.
- 형용사 어휘는 중요 공유시어에 단 두 개만 포함되어 있다.(깊다 00vx, 없다 01vx) 전체적으로 형용사 어휘의 빈도가 적지만 묘사대상의 구체화와 관련된 어휘가 적은 것은 특기할 만하다.

소월과 다쿠보쿠의 중요 공유시어에 나타난 계량적 현상을 분석한 결과, 전반적으로 시적 주체의 적극적 노출 가운데도 시적 자아의 활동과 관련된 서술이 많으며, 시의 장식과 묘사 대상의 구체화 부분에 관해서는 크게 관심을 두지 않은 것으로 보인다.

다. 소월-다쿠보쿠의 차이성 검토

공유시어를 분석하여 시적 어휘의 유사성을 찾아내는 것과는 반대로, 이번에는 시적 어휘의 차이성을 탐색하기 위해서 공유시어가 아닌 시어를 찾아보기로 한다. 사실 시적 어휘의 개별성을 탐구하기 위해서 필자는 개인시어(poetic idiolect)의 개념을 제시한 바 있고, 이를 중심으로 시인의 특성을 파악한 바 있다.[5] 그러나 개인시어는 전체 현대시 코퍼스와의 대비를 통해서 설정하는 것인 반면, 이와 같이 양자 대비의 경우에는 상호적 미사용 시어 중에서 고빈도 시어를 중심으로 검토하는 것이 타당하다고 본다.
상호적 미사용 시어는 앞에서 말한 비공유시어(non-shared word)에 해당한다. 시어의 종수로 본다면 소월은 1,373 종(2,152 종-779 종), 다쿠보쿠는 3,185 종(3,964 종-779 종)이 이에 해당한다. 시어의 갯수로 본다면 소월은 2,707 개(7,949 개-5,242 개), 다쿠보쿠는 8,044 개 (18,936 개-10,892 개)가 된다. 이 중에서 빈도가 높은 시어만을 제시하기로 한다. 소월의 경우는 빈도 8 이상인 시어를, 다쿠보쿠의 경우는 빈도 21 이상의 시어를 제시한다.

표 6 소월-다쿠보쿠 비공유 시어

소월		다쿠보쿠	
시어	빈도	시어	빈도
모르다 00vv	46	아이 01ng	79
그대 00np	42	처럼 wx	74
못 04ma	35	면서 wx	64
임 01ng	31	슬프다 vx	57
저 03np	23	파도 ng	53
줄 04nb	22	향 ng	50
아니 01ma	22	(미상)un	44
오 02ic	21	단지 ma	44
희다 00vx	20	등 05nb	42
다 03ma	19	창문 ng	40
전등 07ng	16	닮다 vv	38
설움 00ng	16	모두 ng	37
리 02nb	15	사라지다 vv	37
듯이 01nb	15	모습 ng	35
두 01mm	15	생 ng	33
못하다 00ⓒvx	14	어둠 ng	33

[5] 김병선(2005), 「석정 시의 계량적 문체 연구 시론」, 석정문학 제 18 집, 석정문학회. pp.140-167 참조.

마음속 00ng	14	환상 ng	33
년 02nb	13	공주 ng	32
냄새 00ng	12	흔적 ng	29
서럽다 00vx	10	가라앉다 vv	27
강물 00ng	10	울림 ng	27
홀로 00ma	9	향내 ng	27
어찌하다 00vv	9	까지 wx	26
삼수갑산 00nm	9	잠시 ma	26
벌 01ng	9	남자 ng	24
달맞이 00ng	9	조용히 ma	23
고요히 00ma	9	풍기다 vv	23
흘러가다 00vv	8	해(年)ng	23
물결 00ng	8	혼 ng	23
동무 01ng	8	여행 ng	22
대로 01nb	8	과연 ma	21

대략적으로 비공유시어를 상호 비교해 보기로 한다.
다쿠보쿠는 유난히 '아이 01ng(79)'에 대한 관심이 많다. 소월은 동요류의 작품에서 시적 화자로 어린이를 등장시킨 경우는 있어도 시적 대상으로 아이를 언급한 적이 없다. 다쿠보쿠는 '공주(公主)'를 시적 대상으로 등장시켰고 '남자(男子)'도 제법 등장한다. 그에 비해 소월의 시적 화자는 대부분 여성 화자로 분석될 수 있으며, '임 01ng'을 줄기차게 부른다.
소월은 한 번도 '슬프다 va' 혹은 '슬퍼하다 vv'라는 표현을 한 적이 없다. 대신 소월은 '설움'과 '서럽다'라는 감정에 사로잡혀 있을 뿐이다. '알다'의 경우는 공유시어로서 소월과 다쿠보쿠 시에 등장하는 비율이 거의 같은 데 비해, '모르다'는 소월만 사용한 시어로 등장한다.[6] 소월의 경우 '알다'가 25 회이고, '모르다'는 46 회로서 모른다는 것의 빈도가 상대적으로 높은 것은 하나의 특징이라 할 만하다. 어쩌면 다쿠보쿠와 달리 소월에게 있어서 세계는 미확정적이고, 불분명하며, 애매한 것으로 가득했는지 모른다.
다쿠보쿠는 '파도' 등 바다와 관련된 시어가 적지 않은데 비해 소월에게서는 '밀물 01ng(6)' 정도만 등장하고, '강물' 등 대부분 강과 관련된 어휘가 많다. 소월은 '냄새'를 맡고 있는 데 비해 다쿠보쿠는 '향'과 '향내'를 탐닉한다. 소월의 시에서는 '혼'은 오로지 시 제목에서만 등장한다.(招魂) 유사한 '영(靈)'은 단 4 회만, '신(神)'은 단 1 회만 등장할 뿐이다. 이에 비해 다쿠보쿠는 '영'도 23 회 등장하며, '신'도 27 회 등장한다. 여기에 '그리스도(Christ)'(2 회)까지 포함한다면 다쿠보쿠는 영적인 세계에 대한 관심이 적지 않았고, 소월은 현실 세계에 대한 관심이 강했음을 확인할 수 있다.
소월의 '그대'는 비공유시어로 되어 있지만, 사실 2 인칭 대명사는 다쿠보쿠에게서도 많이 찾아볼 수 있으므로 이것은 논외로 하는 것이 좋겠다. 3 인칭인 '저'의 경우도 마찬가지다. 그 외에도 조사 등 몇 가지 형태와 관련된 어휘들도 언어적 차이에서 비롯된 것으로 보아 비공유시어에서 다루지 않겠다.

---

6 일본어에는 '모르다'라는 뜻의 어휘가 독립적으로 존재하지 않고, '알다'에 부정사를 덧붙여서 표기한다. 다쿠보쿠 텍스트에서는 '알다'가 총 60회 등장하는데 그 중 부정사가 붙은 것은 7개에 불과하다.

5. 한-중 시어 비교 연구

가. 서지마 텍스트의 확정

현재 서지마 시작품의 발표지면과 원본 시집을 구하기는 매우 어려운 상황이다. 다행히 그 동안 출판된 『서지마 전집』이 5 종이나 전하고 있어서 이를 바탕으로 온전한 서지마 전집의 교정본 텍스트를 재구성하는 것은 그리 어렵지 않다. 다섯 가지 판본을 세밀하게 살펴본 결과, 각각의 편집 의도를 확인하였는데, 전반적으로 그 특징을 요약하면 다음과 같다.

표 7 서지마 전집의 판본 특성

연도	판본명	편집인	작품수	판본특성
1976	대만판	오복출판사	104	번체로 되어 있다. 문장부호가 현재와 많이 다르고 가로로 편집되어 있다. 수록하지 못한 시가 많다.
1983	상해판	상무인서관	104	번체로 되어 있다. 문장부호가 현재와 많이 다르고 가로로 편집되어 있다. 대만판보다 수록 작품이 많아졌지만, 일부 작품은 누락되었다.
1987	절강판	고영제	180	간체로 되어 있다. 문장부호는 현재 중국에서 통용되는 것을 적용하였다. 광서판보다 수록 작품수가 많으나 시 편집에 오류가 매우 많다.
1991	광서판	조하추	177	간체로 되어 있다. 문장부호는 현재 중국에서 통용되는 것을 적용하였다. 시 편집에 있어 많은 오류가 많으나, 수록 작품은 상해판보다 많다.
2005	천진판	한석산	184	간체로 되어 있다. 문장부호는 현재 중국에서 통용되는 것을 적용하려 했다. 시 편집에 있어 엄밀한 태도를 취했다. 수록한 시작품의 정확도가 매우 높은 편이다.

이 연구에서는 천진판의 184 수와 미수록된 <중추월(仲秋月)>과 <칠율(七律)>을 더하여 186 수를 결정본으로 확정하기로 한다.

나. 서지마 텍스트의 형식적 특성

전산 입력한 서지마 전집의 텍스트에 약간의 처리를 한 다음에 형식적 특성을 살펴보았다. 여기서 말한 약간의 처리는 완전한 형식의 용례색인이 아니라 단순한 형태의 어휘 목록이다. 이를 위해서 어휘별 띄어쓰기를 실시하였고, 각 어절에서 문장부호를 제거하였다. 각 어절에서 조사 중 빈도가 높은 조사 '的' 하나만 제거하여 이를 바탕으로 통계처리를 하였다. 그 결과 서지마 시작품의 형식상의 특성은 다음과 같았다.

표 8 서지마-소월 일반 통계

	작품수	어휘	어종	어휘/어종 반복지수	음절	행	작품당 어휘 수
서지마	186 수	32,503 개	9,934 종	3.272	50,040 자	5,255 행	174.75 개
소월	145 수	7,949 개	2,152 종	3.694	n/a	n/a	54.82 개

서지마의 어휘/어종 반복지수는 소월보다 많이 낮은 편이었다. 그의 작품수가 소월보다 훨씬 많고 어휘와 어종도 훨씬 많은 것에 비해 반복지수가 낮은 것은 주목할 만하다. 향후 조사에 대한 처리가 완결된다면 반복지수는 이보다 높아지리라 생각한다. 아직 활용형에서 기본형 추출 작업이 완성된 것은 아니기 때문에 어휘/어종 반복지수에 대한 확정적인 해석은 유보하기로 한다. 다만 작품당 어휘수를 살펴보면, 서지마의 작품들이 평균적으로 3.19 배나 높다는 것이 특징으로 드러난다. 서지마 작품의 길이가 매우 길다는 말이 되겠다.

다. 서지마 시어의 특성

간략한 어휘 목록에 대해서 통계 처리를 한 결과는 다음과 같다.

표 9 서지마 고빈도어 목록

시어	빈도	시어	빈도	시어	빈도	시어	빈도
我	1307	的	61	唉浩	33	将	24
一	840	没有	60	因为	33	快	24
不	795	真	59	太阳	33	片	24
在	747	给	57	边	33	上帝	23
是	712	得	57	时	33	永远	23
你	710	人	56	如	32	天上	23
了	521	都	55	已经	31	黄昏	23
这	446	叫	51	对	31	无	23
有	334	生命	50	变	31	条	23
他	275	心	50	眼	31	做	23
那	261	向	50	先生	30	海	23
个	225	什么	49	人生	30	明星	22
她	225	之	49	起	30	自由	22
也	211	好	49	用	30	只是	22
与	174	过	48	风	30	黑夜	22
里	166	我们	47	比	29	却	22
来	151	浩唉	46	女郎	28	头	22
像	134	可	46	颗	28	尽	22
上	127	多	46	飞	28	阵	22
又	125	死	46	让	28	最	22
说	116	从	46	亦	28	吃	22
似	114	呀	45	已	28	哪里	21
爱	111	人间	43	仿佛	27	青年	21
看	107	啊	43	呢	27	现在	21
去	105	你们	42	成	27	家	21

但	105	问	42	笑	27	住	21
再	99	点	42	如今	26	知	21
谁	96	前	41	梦	26	回	21
就	96	灵魂	40	早	26	可怜	20
能	91	它	40	曾	26	朋友	20
到	91	手	39	着	26	世界	20
要	89	把	39	话	26	深夜	20
只	85	为什么	38	恋爱	25	颜色	20
天	83	知道	38	宇宙	25	曾经	20
见	81	声	38	自己	25	还有	20
中	77	光明	37	星	25	没	20
听	75	走	37	夜	25	二	20
更	73	想	36	座	25	全	20
还	69	间	35	有时	24	正	20
为	66	吧	35	脸	24	和	20
他们	65	下	35	年	24	花	20

서지마의 시어에서 나타나는 몇 가지 특징을 요약하면 다음과 같다.
- 최고 빈도 어휘는 '나(我)'이며, 그 빈도는 제 2 위와도 현격하게 차이가 난다. 이는 앞의 소월이나 다쿠보쿠의 시에서도 확인한 사항이며, 한국현대시에서도 공통적인 현상이다. 말하자면 일반적으로 서지마 시 역시 서정시의 본령에 가깝다고 할 수 있겠다.
- 제 3 위 고빈도 시어는 '아니(不)'라는 부사어이다. 이는 작품 내의 문맥을 살펴서 확인해야 하겠지만 작품의 정서적 경향과도 관련이 있을 것으로 보인다.
- 고빈도어에 속하는 많은 어휘들은 의미와 관련된 어휘들이기보다 언어의 형식에 관련된 어휘들이 많은 것을 알 수 있다.

6. 남는 문제와 해결 전망

이 연구에서는 근대문학 시기의 한중일 삼국의 대표적 근대시인의 작품 세계를 비교 연구하되, 그 어휘에서 나타나는 계량적 현상에 주목하여 상호 비교하는 여러 가지 방법을 논의해 보았다. 이 연구에서는 특히 시작품 코퍼스를 바탕으로 이에 대한 적절한 처리를 통해서 계량적 현상을 밝히고 이를 통계적으로 비교하여 설명하는 방안을 검토해 보았다. 여러 가지로 많은 노력을 기울였지만, 특히 한중일 삼국 언어의 차이에서 오는 여러 가지 문제들이 이와 같은 비교 연구에 많은 장애가 되었다. 한국과는 정서법 면에서 띄어쓰기가 매우 다르기 때문에 중국과 일본의 텍스트 처리에 있어서는 여러 가지 문제가 있었고, 이 연구에서는 이러한 띄어쓰기의 실제에 대해 논의하였다.
띄어쓰기에 이어 중간언어로 설정한 한국어로 번역하는 절차에 있어서 문맥 중심의 번역을 택할 것인가 아니면 일대일로 대응하는 번역을 강구할 것인가에 대해서 논의하였다. 이 연구에서는 우선 일대일 대응에 의한 번역을 실시하였는데, 향후 세밀하게 각 어휘를 대응시켜 번역하는 방안을 강구할 필요가 있다고 보았다. 이를 위해서는 표준적인 번역어 목록을 만들어서 자동 번역에 활용하는 방안도 강구해야 한다.

한편 몇 가지 계량적 처리를 통해서 비교 대상에 대한 설명을 시도하였는데, 여러 가지 계산한 결과 수치에 대한 정밀한 설명이 아직 미흡한 상태이다. 사소한 숫자적 차이가 의미상에서 어떤 변화를 가져오는지에 대해 정밀하게 연구할 수 있어야 한다. 또한 각 어휘에 대해 현재는 문맥에서 독립적으로 검토하였는데, 앞으로 이 어휘들이 실제 시작품 문맥에서 어떤 의미를 가지고 있는지를 충실하게 설명해야 할 것이다.
한편 단지 어떤 어휘가 어떻게 쓰이고 있다는 것에 대한 설명만으로는 대상 시인의 시적 경향이나 시세계를 밝히는 데 역부족일 수 있다. 이런 때는 어휘를 군집하여 비슷한 의미적 범주 혹은 기능적 범주에 따라서 재분류할 필요가 있다. 필자의 분류안은 하나의 방안이 될 수 있다.
이제 어휘에 대한 계량적 비교를 통한 한중일 삼국 시문학에 대한 비교 연구는 앞으로 학위논문을 통해서 보다 구체적으로 발전될 것이다. 이러한 연구가 축적되어 갈수록 한중일 삼국 문학의 관련성은 보다 잘 밝혀질 것이라고 믿는다.

참고 문헌

김병선(2000b),「현대시의 계량적 문체 연구 시론 –문학은 계산될 수 있는가?–」, 『전국학술발표대회 발표논문』, 국어문학회: pp.10-27.
김병선(2001), 『한국 현대시어 용례사전(CD-ROM)』, 누리미디어 KRPIA.
김병선(2004),「한국 현대시 데이터베이스의 구성과 그 활용방안」, 『한국언어문학』 제 53 집, 한국언어문학회. pp.513-535.
김병선(2005),「석정 시의 계량적 문체 연구 시론」, 『석정문학』 제 18 집, 석정문학회. pp.140-167.
김병선(2006),「현대시인의 문체적 지문을 찾아서」, 『국어국문학』 제 143 호, 국어국문학회. pp.153- 188.
김병선(2007),「시적 유사성 탐구 방안 연구」, 『조선-한국학 국제학술대회발표논문집』, 중국 연변대 아세아연구소.
김병선 외(2007), 『한국 현대시어 빈도 사전』, 한국문화사.
김병선(2009a),「詩歌 類似性 探究 方案 硏究」, 『朝鮮-韓國文學與東亞』(延邊大學亞洲硏究中心學術叢書 제 4 집), 中國: 延邊大學出版社. pp.371-400.
김병선(2009b),「시어의 기본형을 찾아서 -목록 참조를 통한 지능적 기본형 추출 방안-」, 『문학 연구와 정보과학 학술회의 자료집』, 한국학중앙연구원 어문생활사연구소. pp.5-28.
김병선(2010),「문체 연구와 코퍼스의 활용」, 『학술회의논문집』, 한국학중앙연구원 어문생활사연구소.
김병선・전정구(1994), 『소월의 시어와 그 쓰임새 1,2,3』, 한국문화사.
만정정(2009), 『김소월과 서지마의 시세계 비교연구』, 석사논문, 단국대학교.
송희복(2008),「김소월과 이시카와 다쿠보쿠의 시세계」, 『한국시학연구』.
오세영(2000), 『김소월, 그 삶과 문학 』, 서울대학교 출판사.
오영진(1994),「이시카와 다쿠보쿠 문학에 나타난 한국관-안중근을 노래한 시를 중심으로-」, 『일본학』.
왕해윤(1974), 『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 硏究』, 석사논문, 경희대학교.
정은혜(2007),「일본근대문학에 미친 한국고전소설의 영향에 대한 고찰 - 메이지(明治) 시대의 나카라이 도스이(半井桃水)와 『구운몽』을 중심으로」, 『학술발표대회논문집』, 한국일본어문학회. pp.243-247.
홍기백(1982),「이시카와 다쿠보쿠 시가와 김소월 시의 비교연구」, 건국대교육대학원 학위논문.