韓國 現代詩의 言語 等級과 文學敎育

金炳善*

1. 문학교육과 언어 등급의 관련성

이 연구는 외국인 학습자를 위한 한국 현대시 작품 선정의 한 방안으로서, 작품에 사용된 어휘에 일정한 등급을 적용하여 작품의 언어적 난이도를 측정하고, 그에 의해 시 작품의 언어적 등급을 평가하는 것을 목적으로 한다. 이 논의는 사실상 문학교육에 있어서 正典(canon) 논의와도 관련을 가진다. 작품 텍스트의 언어적 실체는 정전 선택에 있어 중요한 요인이며, 특히 교육과정의 설계에 있어서 언어적 특성에 대한 고려는 매우 필요한 일이다.

일반적으로 해외의 한국학과에서는 일반적으로 고급한국어 습득과 한국문학에 대한 이해 자료로서 한국 현대시 텍스트를 다루고 있다. 외국어로서의 한국문학교육의 목표는 고급 한국어 교육과 한국문화 교육 및 문학 자체에 대한 감상 능력의 신장을 위한 것으로 요약될 수 있다. 이 중 고급 한국어 교육을 위한 문학 텍스트 선정에 관한 논의는 매우 중요함에도 불구하고 아직 학계에서 본격적으로 다루어지고 있지 않다.

문학작품 중에도 소설이나 수필 장르와는 달리 현대시는 고도의 언어적 조직으로 되어 있어서 작품의 의미 이해 이전에, 작품의 언어 자체를 이해하기 어렵게 만든다. 현대시에는 현대의 표준어뿐 아니라, 방언, 고어 및 개인의 조어까지도 포함되어 있고, 표현하는 방식 또한 비유나 역설과 같은 각종의 문학적 기법이 적용되어 있어서 외국인 학습자는 물론 한국

* 韓國學中央研究院國語敎育學系敎授

의 학생들도 작품 이해가 쉽지 않은 것이 현실이다.

해외 한국학의 한국문학 교육 교재 편찬에 있어 적용되는 현대시 텍스트 선정 기준은 위와 같은 언어적 난이도가 아니라 대체로 문학사적인 의의가 있는 유명작품 위주로 선정되고 있다. 한국어 학습을 고려하거나 문학 기법의 학습을 위한 체계성 등은 크게 고려되고 있지 않다. 실제에 있어서 외국인 학습자를 대상으로 하는 텍스트 선정에 있어서는 한국의 중등학교 교과서에 수록된 작품이 선호되고 있을 뿐이고, 교재 편찬자의 관심에 따른 임의적인 선택 방식이 작용하고 있다고 본다.

이 연구는 외국인 학습자의 한국문학 교육에 사용할 문학작품의 객관적이고 체계적인 선정 방법에 대해 논의하려 한다. 특히 언어적인 난이도의 면에 주목하여, 어휘의 등급 체계 설정 방안을 먼저 연구하고, 이어 어휘의 난이도에 따라 어떻게 작품의 언어적 등급을 부여할 수 있을 것인가를 검토하려 한다. 이를 위해 컴퓨터를 활용한 계량적인 연구 방법을 적용할 것이다. 연구 대상은 필자가 확보하고 있는 61만 어절 규모의 한국현대시 코퍼스이다. 이 코퍼스의 수록 어휘에 대해 등급을 부여하고, 그 결과로 작품의 언어적 등급을 산출하며, 몇 가지 두드러진 통계적 현상에 대해 해석하고자 한다.

2. 어휘 등급 체계의 결정

2.1. 국립국어원의 한국어 학습용 어휘등급 준용

국립국어원의 어휘 등급 연구는 2000년부터 이 기관이 한국어 교육에 필요한 기초 연구를 추진하는 가운데, 2003년에 최종 보고서가 나오면서 마무리 된 것이다. 이를 위해서 2000년부터 2002년까지 한국어 빈도 조사를 먼저 실시하고, 고빈도어 중 상위 10,352개(출현 빈도 15회 이상)를 대상으로 하여, 국립국어원에서 위촉한 선정 위원들(이 위원들은 주요 한국어 교육기관에서 한국어 교육을 담당하고 있는 강사들임)이 등급 판정을 하였다.

이어 몇 차례의 조정 과정을 거쳐서 최종적으로 약 6,000개의 어휘에 대하여 A-B-C 세 등급의 판정을 하였다. 이 A-B-C 등급의 어휘는 한국어능력시험(KPT)에서 상정하는 단계에 준해 각각 1단계-2단계-3단계에 대응하는 것으로 하였다 한다. 최종적으로 선정된 한국어 학습용 어휘는 5,965개이고, 각 등급별로 품사를 구분하여 표를 보이면 다음과 같다

.[1]

표 1 한국어 학습용 어휘 목록 품사별 분포와 비율

품사 \ 등급	A	B	C	계	비율 (%)	KoPoCo 빈도(개)	KoPoCo 비율(%)
명사	497	1,199	1,708	3,404	57.07	250996	41.03
고유명사	21	27	15	63	1.06	7106	1.16
의존명사	33	44	53	130	2.18	16846	2.75
대명사	32	5	10	47	0.79	32914	5.38
수사	45	2	-	47	0.79	1748	0.29
동사	155	501	689	1,345	22.55	149192	24.39
형용사	75	132	169	376	6.30	58808	9.61
보조용언	18	5	10	33	0.55	15534(vx) 2253(vz)	2.91
관형사	27	19	23	69	1.16	23607	3.86
부사	65	146	182	393	6.59	45471	7.43
감탄사	12	22	10	44	0.74	4424	0.72
분석 불능	2	9	3	14	1.06	482	0.08
계	982	2,111	2,872	5,965	100.00	611809	100.00
기타						2428[2]	0.40

한국어 어휘 빈도 조사는 여러 가지 목적으로 몇 차례 이루어진 바 있고, 개인적 차원에서 한국어 교육용 어휘 선정도 시도된 바 있지만, 공적인 기관의 한국어 학습용 어휘 선정 작업은 이것이 처음이라고 할 수 있다. 비록 이 조사가 학문적이고 이론적 접근 방식이 아니라 한국어 교수-학습의 현장에 있는 연구자의 경험적이고 주관적인 접근 방식을 취했다는 점에서 향후 개선의 여지가 있겠으나, 국립국어원 측에서 가급적 객관적 타당성을 확보하고자 많은 노력을 기울인 연구 결과이며, 논의의 시작을 위한 중요한 기반을 마련한 것이라고 평가할 수 있겠다. 따라서 이 연구에서는 이 등급 체계를 받아들여 현대시 텍스트의 언어 등급 부여 방안을

1 조남호(2003), 한국어 학습용 어휘 선정 결과 보고서, 국립국어원. p.11.에서 표를 인용하고 여기에 필자가 비율을 덧붙였다.

2 KoPoCo 빈도에서 기타에 속하는 것은 어미 3, 조사 1836, 접두사 77, 접미사 88, 어근 424 등이다.

강구해 보려 한다.

2.2. KoPoCo의 어휘 등급 처리와 분석

필자는 1980년대 후반부터 한국현대시 데이터베이스를 구축하여, 이를 바탕으로 『한국현대시어 용례사전』[3]과 『한국현대시어 빈도사전』[4]을 편찬 출간한 바 있다. 이 데이터베이스에는 1923년부터 1950년 사이에 출판된 한국 현대 창작시집의 창작시 작품 8,201편(시인의 총수는 344명)의 원문과 용례 및 각 어휘에 대한 언어학적 분석 정보가 수록되어 있다.[5]

이 데이터베이스는 지속적으로 보완, 수정의 과정을 거쳤으며, KoPoCo는 이 데이터베이스의 별칭(Korean Poetry Corpus)이다. 계속적인 정제 작업을 거치면서 각 연도 숫자를 붙여서 구분하고 있다.(현재는 KoPoCo-2011) KoPoCo-2011의 어휘의 총수는 611,809개이고, 어종의 총수는 42,453종(정상화를 거쳐 관련어를 기준으로 한 경우 34,981종)이다.

시작품의 언어적 등급을 부여하기 위해서 먼저 KoPoCo의 어휘에 대하여 등급을 부여하고, 부여된 어휘 등급을 재처리하여 작품별 등급을 정하기로 한다. 작품의 등급이 정해지면 작가별 혹은 장르별 언어적 등급의 제반 현상에 대해서 검토할 수 있게 된다.

2.2.1. 어휘 표기의 정상화

KoPoCo의 어휘 표기 방식과 『보고서』의 방식은 약간 차이가 있다. 따라서 목록의 상호 대조를 통한 자동 처리를 위해서는 표기 방식을 통일해야 한다.

먼저 국립국어원 『보고서』는 '단어'를 표시할 때, 동음이의어가 있는 어휘에는 두 자릿수의 아라비아 숫자를 첨자(superscript)로 붙였고, 동음이의어가 없는 어휘에는 아무런 표시를 붙이지 않았다.(예: 가족01, 가지01, 가지04, 가지다, 가짜) 이 첨자는 『표준사전』 데이터베이스의 번호와 동

---

[3] 김병선 외(2001), 『한국 현대시어 용례사전』, 누리미디어 KRPIA.
[4] 김병선 외(2007), 『한국 현대시어 빈도사전』, 한국문화사.
[5] KoPoCo의 자세한 구성에 대해서는 다음 논문을 참고하시오. 김병선(2010), 문체 연구와 코퍼스의 활용, 『차세대 어문정보학의 전망 학술회의 논문집』, 한국학중앙연구원 어문생활사연구소. pp.33-55.

일하며, 인터넷 사전(stdweb2.korean.go.kr)의 번호와 같다.[6] 품사 이름은 별도의 필드로 제시하였고, 어휘의 의미 이해를 돕기 위해 한자나 외래어 및 간략한 용례 등을 '풀이'로 제시하였다.

KoPoCo 데이터베이스 체계는 '어휘-첨자-품사기호' 형식의 문자열로 표시되기 때문에 다음과 같이 수정하였다. 먼저 동음이의어가 없는 어휘에 '00' 번호를 붙였고, 품사명은 필자가 사용하고 있는 두 자릿수의 영문 표기로 바꾸어서, 최종적으로 어휘와 첨자와 품사기호를 묶었다. 품사기호는 '21세기 세종계획'의 기호를 반영하되, 필자가 간략화한 영문 알파벳 두 글자로 통일하였다. 세종계획과는 달리 보조동사, 보조형용사는 구분하였다.[7]

2.2.2. 어휘 품사의 확장

국립국어원의 보고서에 따르면 관형사와 수사의 표기가 동일한 경우는(특히 한자어의 경우) 관형사는 제외하고 수사만 목록에 남겼다고 한다.[8] 사실 '일, 이, 삼, 사'처럼 수사와 관형사가 동일한 경우에는 이를 '수관형사' 혹은 '수사'로만 처리할 수도 있으나, KoPoCo에서는 이를 문맥에 따라 수사와 관형사로 구분해 두었다. 따라서 이번 연구에서는 비록 보고서에 등재되지 않은 관형사라도, 동일 형태로 수사에 포함되어 있으면, 같은 등급의 관형사로 처리하였다.

또한 '감정적01', '결과적00'처럼 '-적(的)'이 붙은 말이 관형사와 명사 등 두 가지 품사로 분석될 수 있을 때에, 보고서에서는 원칙적으로 명사로만 취급하였다. 이번 연구에서는 수사와 관형사의 경우와 마찬가지로 같은 등급의 관형사와 명사를 모두 목록에 포함하는 것으로 하였다. 이렇게 하여 당초 보고서 목록에는 69종의 관형사만 등록되어 있었으나, 이번 연구에서는 KoPoCo의 여러 사례를 반영하여 모두 82종의 관형사가 C등급 이상으로 처리되었다.

---

[6] 인쇄본, CD-ROM 판 및 <아래한글>에 포함되어 있던 국어사전의 첨자 정보와는 다소 다르다.

[7] 일반명사(ng), 고유명사(nm), 의존명사(nb), 대명사(np), 수사(nr), 동사(vv), 형용사(va), 보조동사(vx), 보조형용사(vz), 관형사(mm), 부사(ma), 감탄사(ic), 격조사(jk), 어미(ed), 접두사(xp), 접미사(xs), 어근(xr), 기호(sb), 분석불능(un)

[8] 조남호 편(2003), 한국어 학습용 어휘 선정 보고서, 국립국어원. pp.9-10.

2.2.3. 관련어 어휘의 적용

분석 대상 작품들이 1950년 이전에 발표된 것이어서 사실상 어휘의 표기가 현재의 표준형 표기와 다른 것이 많다. 게다가 시인이 의도적인 표기도 있어서 KoPoCo에서는 원칙적으로 시인의 표기를 존중하여 기본형을 설정하되, 명백한 오표기에 대해서만 정표기를 제시하였으며, 옛한글의 경우는 대응되는 현대 한글로 치환하는 등의 정상화(normalization) 처리를 하였다.

이러한 처리를 거친 기본형 어휘는 현대 맞춤법에는 어긋나는 것이 적지 않기 때문에, 작품의 의미 분석 관련 처리에는 장애가 되었다. 따라서 작품 표기에 근접한 기본형에 대응하는 현대 표준어를 별도로 제시하였는데, 이를 '관련어'로 부른다. 국립국어원의 『보고서』가 현대표준어에 대해 등급을 부여한 것이기 때문에, 이번 연구에서는 KoPoCo의 관련어를 적용하여 처리하기로 하였다.[9]

몇 가지 예를 들면, 현대시 코퍼스에서는 시인의 표기에 따라 '님05'(1167회)과 '임01'(512회)을 모두 기본형으로 설정하였다. '님05'은 '임01'의 옛말인데, 기본형으로서는 각각 다르게 되어 있지만, 관련어로서는 '님05'에 대해 '임01'만 표시하였다. 결과적으로 '님05'에 대해 '임01'과 동일 등급을 부여한 셈이다. 또한 『표준사전』에 의하면 '기대이다00'는 '기대다01'의 북한어로 풀이되어 있는데, 사실상은 '기대다01vv'의 잘못이다. '우05'는 '위01'의 잘못으로, '칩다00'는 '춥다00'의 방언으로 풀이되어 있다. 이러한 오류 어휘에 대해서는 모두 바른 표기의 관련어를 기본형으로 처리하여 등급을 부여하였다. '아즉90'처럼 『표준사전』에는 등재되지 않은 어휘의 경우도 바른 표기인 '아직01'을 관련어로 처리하였다. '저편00'은 '저쪽00'의, '소낙비00'는 '소나기01'와 동의어인데, 이런 동의어의 경우도 후자의 표기를 관련어로 처리하여 등급을 부여하였다.

한편 '맘01'은 '마음00'의 준말이므로 KoPoCo에서는 '맘01'에 대하여 '마음00'을 관련어로 처리하였다. 그런데 국립국어원의 『보고서』에서는 '

[9] 이것은 향후 한국문학 교재 편찬에 있어서 텍스트를 어떻게 확정할 것인가의 정책 문제와도 관련이 된다. 말하자면 텍스트의 표기를 현대어로 하느냐 아니면 당대의 표기대로 하느냐, 표준어로 하느냐 아니면 원본대로 하느냐, 나아가서는 다양한 판본 중 어떤 판본을 저본으로 할 것인가, 또 원전비평을 어떤 수준으로 할 것인가의 문제와 직결된다. 일단 이 연구에서는 오표기, 잘못, 방언(대응되는 표준어가 있는 방언) 등을 현대 표준어 정표기로 수정하는 것을 전제로 한다.

맘01'에는 C등급을, '마음00'에는 A등급을 각각 부여하고 있다. KoPoCo에서는 『보고서』를 존중하여 각각 다른 등급을 부여하였다.

2.2.4. 5등급 체계로의 확장

국립국어원의 『보고서』는 현대국어 고빈도어 약 6천 개를 대상으로 이를 3등급 체계로 분류하고 있다. 이 대상 목록에 속하지 않는 어휘는 '등급외'로 분류할 수 있으므로, 사실상은 4등급 체계인 셈이다. 『보고서』의 등급외 어휘(즉 제 4등급)는 표준어인데 비해, KoPoCo의 경우 등급외 어휘가 『표준사전』에 수록된 어휘(표준어와 비표준어를 포함함.)와 미수록 어휘로 구분될 수 있다. 물론 『표준사전』에 수록되었더라도 표준어가 아닌 지역어 혹은 오류어일 수도 있으나, 이 연구에서는 등급외 어휘 중 『표준사전』 수록 표제어는 D등급으로, 미수록 표제어는 E등급으로 처리하기로 한다. 또한 원본의 인쇄 불량 등으로 기본형을 제대로 설정할 수 없는 어휘에 대해서는 F등급을 적용하며 실제 처리에서는 제외하기로 한다.

3. KoPoCo 어휘 등급 부여

3.1. 어휘별 등급의 일반적 특징

이와 같은 사전 조정을 거쳐 이 연구에서는 61만 어절 규모의 KoPoCo 어휘에 대해 각각 등급을 부여하였다.[10] KoPoCo는 MS Access의 테이블 형태로 관리되고 있는데, 그 중심 테이블인 concordance 테이블에 등급 필드를 추가하고, 국립국어원의 어휘등급표 테이블과 연결하였으며, join update 쿼리를 실행하여 자동으로 각 어휘 레코드에 어휘 등급을 부여하였다. 그리고 향후 활용을 위하여 KoPoCo의 장르 분류도 반영하였다. 이러한 처리 결과를 요약하여 각 작품별로 등급 현황을 하나의 테이블로 집계하였다. 각종 계산 처리는 MS Access 테이블을 MS Excel 시트로 옮겨서 처리하였다. KoPoCo-2011에 대한 어휘 등급 처리 결과는 다음과 같다.

표 2 KoPoCo-2011의 어휘 등급 처리 결과

등급	어휘(token)	어종(type)	어휘/어종	어휘비율	어종비율
A	262,197개	874종	300.00	42.86%	2.50%
B	118,127개	1,662종	71.08	19.31%	4.75%

[10] 이번 연구에서는 작품의 제목은 제외하고 시 본문에 대해서만 평가한다.

C	67,072개	1,872종	35.83	10.96%	5.35%
D	154,050개	24,020종	6.41	25.18%	68.67%
E	9,799개	6,403종	1.53	1.60%	18.30%
F	564개	150종	3.76	0.09%	0.43%
합계	611,809개	34,981종	17.49	100.00%	100.00%

사실 필자의 관심은 D등급 어휘의 빈도와 반복지수에 있었다. 왜냐하면 일반어를 대상으로 하는 빈도조사를 바탕으로 한 국립국어원의 등급 체계가 문학어 특히 시어에 적용할 때에 무리가 없을까 하는 것을 확인할 수 있는 시금석이기 때문이다. 일단 어휘의 반복지수(어휘/어종) 면에서 볼 때에 A등급이 가장 높고, 밑으로 내려갈수록 현격한 차이로 낮아지는 것을 확인할 수 있다. 그런데 A등급의 반복지수가 300.00에 이르는데 비하여 D등급은 반복지수가 6.41에 불과한 것으로 나타났다. 이처럼 D등급으로 분류된 어휘들이 시작품에서도 그리 높은 빈도를 보이지 않았고, 각 등급별로 반복지수가 일정한 패턴을 지니고 있으므로, 이 등급체계는 현대시 텍스트에 적용하는 데에도 어휘빈도 면에서는 타당성을 가지고 있는 것이라는 것을 말해 준다.[11]

A등급 어휘(token)의 비율이 42.86%에 달하고, B등급까지 합하면 60%를 상회하는 것을 본다면, 일반적으로 어휘적인 면에서 현대시가 아주 난해한 것은 아니라고 일단 평가할 수 있다.[12]

또 한 가지 특징으로 KoPoCo 고빈도어들이 A 등급을 받은 것을 알 수 있다. 즉 대체적으로 KoPoCo의 고빈도어는 현대국어에서도 고빈도에 속하는 것이다. 그런데 KoPoCo 시어 중에서 국립국어원의 학습용 어휘에 포함되지 않는 몇 가지 어휘가 나타났고, 등급이 낮은 어휘 중에도 고빈도 현대시어가 발견되었다. 이러한 어휘들은 일반어와 시어의 차이점을 말해 주는 것이라 할 수 있다. 그 목록은 다음과 같다.

표 3 KoPoCo 고빈도어의 등급 부여 현황

어휘	국립국어원 등급	KoPoCo 빈도(순위)
소리01ng	B	2838(15위)

[11] 그럼에도 불구하고 향후 24020종에 달하는 D 등급에 대해서는 문학어라는 점에서 재조정할 필요가 있다고 생각한다.

[12] 이후 등급에 대한 논의에서 A등급 쪽으로 갈수록 높고, E등급으로 갈수록 낮다고 표현하기로 한다.

그대00np	C	2002(28위)
임01ng	등급 외	1679(34위)
푸르다00va	B	1624(38위)
태양02ng	B	1446(45위)
흐르다01vv	B	1324(52위)
당신02np	B	1187(65위)
없이00ma	B	1185(66위)
오02ic	B	1082(74위)
희다00va	B	1070(76위)
피다01vv	B	1068(77위)
땅01ng	B	1003(81위)
피02ng	B	913(93위)
붉다01㉠va	B	854(97위)
줄04nb	B	838(100위)
아아01ic	B	822(102위)
그립다00va	B	783(106위)
곱다02va	B	777(107위)
물결00ng	C	759(112위)
깊다00va	B	683(124위)

KoPoCo에서는 고빈도어임에도 등급이 낮은 어휘는 대체로 다음과 같은 특성을 가지고 있다. 먼저 '그대', '당신' 등은 이인칭 대명사로서 시작품에서 호칭어로 많이 사용되는 말이다. '임'의 경우는 대명사는 아니지만 기능상으로는 시작품에서 호칭어로 사용된다. 일반 텍스트에서는 거의 청자에 대한 호칭이 거의 없으므로 '등급외'로 분류되었던 것으로 보인다.

'푸르다', '희다', '붉다' 등의 색채어 형용사나, '그립다', '곱다', '깊다' 등의 상태를 나타내는 형용사가 시작품에서는 고빈도에 속하나 일반 텍스트에서는 B등급으로 분류되고 있다. 명사어 중에서는 '소리', '태양', '땅', '피', '물결' 등의 어휘가 KoPoCo보다 낮은 등급을 받은 것으로 볼 수 있다. '아아' 같은 감탄사가 일반 텍스트보다 시에서 높은 빈도를 보이는 것은 당연한 일이나, 의존명사 '-줄'과 같은 허사류 어휘가 KoPoCo의 고빈도어가 되는 것은 특기할 만한 일이다.

3.2. 분석 대상 작가와 작품의 범위 한정

KoPoCo 데이터베이스에 한 작품이라도 수록된 사람은 모두 344명인데, 이들 중에는 작품의 양도 적고, 이후 별다른 활동을 하지 않아 문학사에서 언급되지 않는 사람도 적지 않다. 따라서 의미 있는 시인의 작품으로 한정하기로 하고, 작품 수가 많은 상위 100위까지의 시인(22편 이상의 작품이 수록됨)을 대상으로 처리하였다.

101편 이상(25명) 시인명(작품수, 어휘수)

金億(343, 20274), 金東煥(268, 41053), 尹崑崗(225, 13621), 柳致環(218, 12818),張貞心(205, 7,299), 金東鳴(191, 10980), 李光洙(182, 8822), 金起林(172, 14067),朱耀翰(158, 11070), 黃錫禹(151, 6803), 盧子泳(146, 10153), 李海文(146, 14109), 金素月(145, 7949), 林學洙(128, 14667), 李殷相(125, 6831), 安自山(117, 2776), 鄭芝溶(113, 7655), 李燦(112, 10039), 異河潤(110, 6269), 毛允淑(109, 8096), 朴鍾和(108, 11698), 趙重治(107, 8802), 金炯元(105, 8681), 李雪舟(103, 4506), 朴貴松(101, 3263)

51-100편 (32명)

權九玄(97, 2442), 權煥(97, 7818), 金達鎭(94, 4049), 盧天命(90, 5060), 金珖燮(89, 6152), 韓龍雲(88, 8245), 尹東柱(88, 3666), 曺雲(85, 2324), 吳章煥(84, 8880), 李庸岳(83, 5966), 李秉岐(83, 3589), 金相沃(75, 5562), 辛夕汀(74, 5318), 朴龍喆(72, 6364), 尹永春(71, 2813), 金永郎(70, 3479), 林和(69, 14692), 李熙昇(67, 3483), 許利福(65, 5575), 沈熏(63, 5826), 張萬榮(60, 4180), 朴八陽(59, 5050), 丁薰(57, 3094), 薛貞植(57, 9590), 金容浩(56, 4565), 金尙勳(55, 11021), 徐廷柱(54, 3974), 金燾星(54, 3619), 梁柱東(53, 3465), 韓竹松(51, 3083), 黃順元(51, 4819), 이태환(51, 3632)

22-50편 (43명)

金大鳳(50, 2934), 金光均(49, 2975), 鄭寅普(48, 5792), 朴世永(48, 6728), 呂尙玄(45, 4161), 趙明熙(42, 3103), 崔南善(38, 2259), 張泳暢(36, 1852), 李相弼(36, 2868), 白石(36, 2219), 金容得(35, 1967), 金相瑗(35, 1185), 鄭昊昇(34, 4256), 朴斗鎭(34, 4168), 金東錫(33, 2166), 卞榮魯(33, 1530), 申瞳集(32, 1952), 朴巨影(31, 4025), 趙芝薰(31, 1830), 趙炳華(31, 1485), 金哲洙(31, 2071), 鄭鎭業(30, 2406), 金春洙(29, 1590), 朴木月(29, 962), 朴文緖(27, 2951), 朴勝杰(27, 1260), 咸允洙(27, 617), 皮千得(27, 1072), 朴魯春(27, 966), 兪鎭五(26, 3710), 金常民(25, 4887), 金明淳(25, 1775), 朴芽枝(25, 3953), 李陸史(24, 1918), 韓何雲(24, 1180), 尹復九(24, 1864), 李箱(24, 2372), 沈仁燮(23, 2134), 李元熙(23, 1279), 林春吉(23, 1310), 申石艸(23, 1736), 金嵐人(22, 1355), 李範赫(22, 1965)

이 100명 시인들의 작품수는 모두 7,494편에 이른다. 이 작품에 나타

나는 모든 어휘에 대해서 A부터 F까지의 등급을 부여하였다. 그리고 각 작품별로 다음과 같은 테이블을 만들어서 등급별 어휘수를 조사하여, 작품별 평균등급을 산출하였다. 그리고 향후 활용을 대비하여 하위 장르를 구분해 두었다.(다음 표에서는 지면 관계로 극시, 서사시 등은 제외함.)

표 4 각 작품별 등급 현황 테이블의 일부

작가	장르	작품제목	자유시	산문시	시조	평균등급	등급A	등급B	등급C	등급D	등급E	등급F
金起林	P	江		49		3.6939	21	8	5	14	1	0
金起林	P	물레방앗간		48		3.5208	18	5	10	14	1	0
金起林	P	北行 列車		40		3.6000	15	8	5	11	0	1
韓何雲	F	개구리	6			1.0000	0	0	0	0	6	0
權九玄	F	부지깽이?	29			1.4138	5	2	2	1	0	19
金東鳴	F	北平	6			1.8333	0	0	1	3	2	0
李光洙	F	뒷동산에	27			2.2593	4	2	3	6	12	0
丁 薰	F	노을	13			2.3077	2	1	0	8	0	2
金相瑗	F	位置	18			2.3333	1	3	1	9	4	0
이태환	F	荊冠	8			2.3750	0	0	3	5	0	0
李雪舟	F	大雄殿	13			2.3846	1	1	0	11	0	0
安自山	K	落花岩			15	2.4000	0	5	0	6	4	0
安自山	K	夢記			17	2.4118	2	1	1	11	2	0
沈 熏	K	南屛 晩鐘			15	2.4667	0	3	2	9	1	0
安自山	K	東京曲			14	2.5000	2	0	1	11	0	0
이태환	F	佛像	8			2.5000	0	2	1	4	1	0
安自山	K	北京			16	2.5000	1	2	2	10	1	0
趙炳華	F	海岸의 설	28			2.5000	3	2	2	20	1	0
金 億	F	넝쿨타령	83			2.5301	14	5	1	54	9	0
朴魯春	F	미꾸리	14			2.5714	0	3	2	9	0	0
金東鳴	F	文字의 悲哀	7			2.5714	1	0	1	5	0	0

위 테이블에서 각 작품별 평균 등급은 다음과 같은 방법으로 계산하였다.

$$\text{작품어휘평균등급}(r) = \frac{\text{어휘점수}(s)_1 + \text{어휘점수}(s)_2 + \text{어휘점수}(s)_3 + \cdots\cdots + \text{어휘점수}(s)_n}{n(\text{어휘 개수})}$$

이를 간단히 수식으로 표현하면 다음과 같다.

$$r = \sum_{1-n} \text{어휘점수}$$

어휘 등급으로부터 작품의 등급을 계산해 내기 위하여, 각 등급 어휘에 대하여 일정한 가중치 점수를 부여하고, 각 어휘점수의 합계를 어휘 총 개수로 나누는 방법을 택했다. 사실 어휘 등급별 가중치를 어떻게 부여하느냐에 따라 작품의 평균 등급은 달라지므로, 그 가중치 부여가 매우 중요하다. 특히 등급별로 점수의 간격을 달리한다든지, 그 간격의 비율이 일정하지 않으면 작품의 평균 어휘 등급은 타당성을 확보할 수 없게 된다.

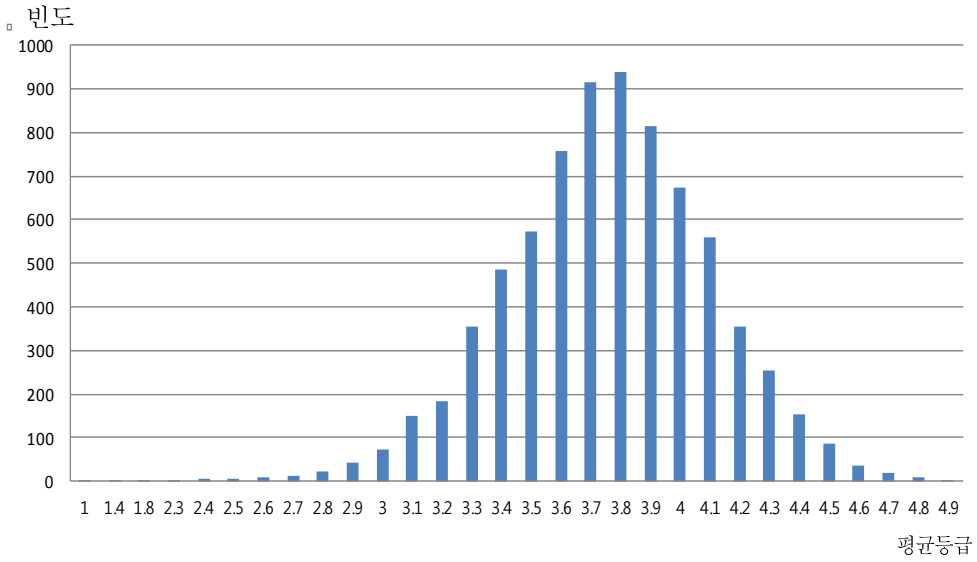
이 연구에서는 일단 어휘 등급별 가중치는 A등급: 5점, B등급: 4점, C등급: 3점, D등급: 2점, E등급: 1점, F등급: 0점으로 처리하였다. 말하자면 등급간 간격을 1점으로 하여 등급 체계의 점수를 부여하였다. 따라서 작품 어휘 등급 계산은, 각 등급 어휘에 가중치 점수를 곱한 다음, 전체 어휘 점수를 모두 합하고 이를 전체 어휘 개수로 나누면 된다. 이렇게 하면 작품어휘등급(r)의 점수는 $1 \leqq r \geqq 5$의 범위에 있게 된다. 이런 식으로 모든 작품에 대해서 어휘등급 점수를 부여하였고, 평균 등급을 소수점 4째 자리까지 계산하였다. 그 결과 다음과 같은 통계치를 얻을 수 있었다.

표 5 작품별 어휘등급 통계

분류	통계 항목	수치
분석 대상	작품수	7,494 편
	어휘수	591,077개
작품별 어휘등급	평균값 average() 표준편차 stdev() 중간값 median() 최빈값 mode() 최대값 max() 최소값 min()	3.7580 0.3470 3.7634 4.0000 4.9167 1.0000

이어서 대체적인 분포 양상을 보기 위해서 각 작품별 평균 등급을 소수점 첫째 자리까지로 재조정하고, 이에 따라 재계산하여 다음과 같은 결과를 얻었다.

그림 1 작품별 어휘등급 분포



평균 등급을 소수점 첫째 자리로 조정하여 계산해 보니 평균값(average)과 중간값(median) 및 최빈값(mode)이 모두 3.8으로 나왔다. 말하자면 이론적으로 표준정상분포(Standard normal distribution)를 보이고 있다고 할 수 있다. 표준편차는 0.3인데, '평균값 표준편차'의 경우는 5,231 작품이 포함되어 모두 69.80%(기준 68% 이상)를, '평균값 표준편차*2'의 경우는 7,017 작품이 포함되어 모두 93.63%(기준 95% 이상)를, '평균값 표준편차*3'의 경우는 7,423 작품이 포함되어 모두 99.05%(기준 99% 이상)를 차지한다. 이러한 정상분포는 그래프 상에서도 평균값을 중심으로 대칭적인 모습을 보인다. 평균 3.0 이하의 경우는 빈도가 매우 낮은 편이므로 이를 제외한다면 보다 더 정상분포에 가까울 것으로 보인다.

4. 시인과 작품의 어휘 등급 분석

4.1. 시인별 통계 분석

이와 같은 처리를 거쳐서 이를 다시 100명의 시인별로 평균을 내보았다. 또한 각 작품 중 등급의 최대값, 최소값 및 표준편차를 살펴보았다.[13]

[13] 시인의 작품별로 통계 처리를 하지 않고 전체 어휘를 대상으로 할 경우, 계산 결과는

표 6 시인별 어휘 등급 통계

번호	시인	작품수	등급 평균	최대값	최소값	표준 편차
1	韓龍雲	88	4.0607	4.5926	3.4194	0.2507
2	朴貴松	101	4.0329	4.7742	3.2000	0.3135
3	朴巨影	31	4.0092	4.4949	3.5692	0.2747
4	朴八陽	59	3.9602	4.6071	3.4130	0.2121
5	皮千得	27	3.9564	4.7222	2.8000	0.3853
6	金素月	145	3.9487	4.6667	3.2083	0.3054
7	盧子泳	146	3.9434	4.6111	3.1563	0.2927
8	異河潤	110	3.9394	4.5909	3.1622	0.2871
9	金珖燮	89	3.9259	4.9167	2.7571	0.3342
10	朴龍喆	72	3.9087	4.4643	3.2857	0.2353
11	朱耀翰	158	3.9079	4.7333	2.8148	0.3514
12	朴世永	48	3.9079	4.3721	3.4438	0.2420
13	張貞心	205	3.8925	4.7895	2.8824	0.3784
14	尹永春	71	3.8918	4.8571	3.2500	0.3365
15	李光洙	182	3.8900	4.6957	2.2593	0.3408
16	金炯元	105	3.8781	4.7692	3.1667	0.2781
17	金明淳	25	3.8731	4.2957	3.4688	0.2226
18	張泳暢	36	3.8723	4.5455	3.3673	0.2769
19	卞榮魯	33	3.8710	4.4286	3.4000	0.2468
20	韓何雲	24	3.8693	4.7241	1.0000	0.6844
21	梁柱東	53	3.8679	4.4795	3.1442	0.2893
22	金大鳳	50	3.8628	4.6154	3.0192	0.3692
23	張萬榮	60	3.8600	4.3772	3.1613	0.2425
24	辛夕汀	74	3.8588	4.4783	3.0606	0.2838
25	權九玄	97	3.8540	4.8000	1.4138	0.4136
26	林　和	69	3.8529	4.2266	3.3175	0.1934
27	李　箱	24	3.8473	4.7193	2.7174	0.4273
28	尹東柱	88	3.8332	4.4906	2.8182	0.3569
29	金東煥	268	3.8299	4.6316	3.0682	0.3017
30	李殷相	125	3.8251	4.5135	3.1212	0.2898
31	朴斗鎭	34	3.8247	4.3788	2.9796	0.2669
32	盧天命	90	3.8225	4.5600	3.2500	0.2612

약간 달라질 수 있다. 이에 따라 순위에 차이가 나오기도 한다. 대표적인 경우가 표준편차가 큰 한하운의 경우다. 이 표에서는 20위이지만 전체 어휘를 대상으로 한 통계에서는 한하운은 3위에 오른다.

33	毛允淑	109	3.8223	4.7308	3.2188	0.2812
34	李範赫	22	3.8171	4.2899	3.4671	0.2278
35	朴勝杰	27	3.8114	4.5455	3.0746	0.3574
36	黃錫禹	151	3.8089	4.4706	2.8462	0.3122
37	崔南善	38	3.8078	4.2941	3.2222	0.2831
38	徐廷柱	54	3.8049	4.5294	3.0577	0.2861
39	薛貞植	57	3.7967	4.4737	3.1667	0.2821
40	金　億	343	3.7963	4.7714	2.5301	0.3602
41	金東錫	33	3.7922	4.4211	3.2821	0.2726
42	金容浩	56	3.7911	4.7500	3.2727	0.2820
43	趙明熙	42	3.7848	4.3846	3.0667	0.3381
44	林春吉	23	3.7821	4.3019	3.3929	0.2607
45	金春洙	29	3.7743	4.7000	3.0476	0.3765
46	金達鎭	94	3.7728	4.5000	3.1053	0.2781
47	兪鎭五	26	3.7695	4.3125	3.3302	0.2271
48	金東鳴	191	3.7513	4.7857	1.8333	0.3889
49	朴文緖	27	3.7500	4.6316	3.2449	0.3019
50	曺　雲	85	3.7483	4.6667	2.6429	0.3987
51	權　煥	97	3.7456	4.3571	2.9333	0.2521
52	李庸岳	83	3.7455	4.4000	2.9800	0.2904
53	李相弼	36	3.7373	4.3036	2.9412	0.3283
54	金永郎	70	3.7348	4.5625	3.2353	0.2653
55	吳章煥	84	3.7245	4.5652	2.6000	0.3596
56	尹崑崗	225	3.7242	4.7000	2.9000	0.3344
57	金哲洙	31	3.7219	4.2600	3.0556	0.3169
58	黃順元	51	3.7179	4.3333	2.6667	0.3128
59	李陸史	24	3.7095	4.0933	3.3333	0.2341
60	金相瑗	35	3.7055	4.5357	2.3333	0.4620
61	李秉岐	83	3.7026	4.4722	3.0625	0.2643
62	金燾星	54	3.7024	4.3571	3.0000	0.3164
63	朴芽枝	25	3.7002	4.2958	3.2157	0.3034
64	趙炳華	31	3.6995	4.5000	2.5000	0.3953
65	申石艸	23	3.6969	4.2364	3.2034	0.2559
66	朴鍾和	108	3.6862	4.2979	2.9216	0.3095
67	鄭寅普	48	3.6856	4.0789	2.9608	0.2338
68	金相沃	75	3.6771	4.6053	3.0909	0.2938
69	鄭鎭業	30	3.6731	4.3158	3.1901	0.2102
70	林學洙	128	3.6717	4.5000	2.7627	0.3063

71	金尙勳	55	3.6715	4.3784	3.2816	0.2186
72	金起林	172	3.6710	4.4186	2.9000	0.2914
73	李海文	146	3.6626	4.5556	2.9798	0.2686
74	柳致環	218	3.6511	4.5000	2.6000	0.3096
75	金光均	49	3.6468	4.4167	3.1967	0.2573
76	趙芝薰	31	3.6462	4.1463	3.0714	0.2634
77	鄭昊昇	34	3.6383	4.0404	3.1098	0.2591
78	申瞳集	32	3.6315	4.0794	3.2128	0.2460
79	李熙昇	67	3.6301	4.3158	2.7000	0.3921
80	沈　熏	63	3.6249	4.3214	2.4667	0.3456
81	鄭芝溶	113	3.6136	4.5000	2.8000	0.3322
82	李元熙	23	3.5980	3.9796	3.0606	0.2435
83	韓竹松	51	3.5848	4.2326	2.7937	0.3055
84	李　燦	112	3.5823	4.4483	3.0735	0.2873
85	白　石	36	3.5806	4.2079	3.0000	0.2792
86	朴魯春	27	3.5600	4.2188	2.5714	0.3702
87	金常民	25	3.5590	4.0000	3.2892	0.1624
88	金嵐人	22	3.5551	3.9500	3.0806	0.2214
89	李雪舟	103	3.5544	4.2593	2.3846	0.3192
90	呂尙玄	45	3.5513	4.4167	2.9583	0.2797
91	尹復九	24	3.5419	3.9750	2.9494	0.2615
92	朴木月	29	3.5383	4.2558	2.8333	0.3781
93	丁　薰	57	3.4978	4.1429	2.3077	0.3508
94	許利福	65	3.4871	4.3000	2.7564	0.3113
95	金容得	35	3.4653	4.6111	2.5938	0.3252
96	咸允洙	27	3.4601	4.1905	2.6818	0.3512
97	沈仁燮	23	3.4552	4.1111	2.9661	0.3291
98	이태환	51	3.4509	3.8788	2.3750	0.2914
99	趙重洽	107	3.4442	4.3488	2.7000	0.3242
100	安自山	117	3.3207	4.3913	2.4000	0.3902

4.2. 작품 등급의 결정 요인 분석

통계 처리의 결과 안자산의 작품이 평균적으로 가장 능급이 낮은 작품으로 밝혀졌다. 그 주된 이유는 장르 특성이라고 할 수 있다. 그의 작품은 모두가 시조로서 아어체(雅語體) 문장의 한자어를 많이 사용하고 있다. 이를테면 ‘가을바람’ 대신 ‘추풍(秋風)’을, ‘내일 아침’ 대신 ‘명조(明朝)’를, ‘하늘과 땅’ 대신 ‘건곤(乾坤)’을, ‘오늘’이나 ‘요사이’ 대신 ‘금일(今日)’을, ‘

매해' 대신 '연년(年年)'을 쓰는 식이다. 아울러서 시조 작품 중에서도 기행시가 많아서 지명, 인명 등 고유명사의 비율(3.6023%)이 매우 높은 것도 등급을 낮추는 데 크게 기여한다. 고유명사어는 『보고서』에서도 단지 64개 어휘만 포함되어 있어서 그 외의 것은 모두 D등급 이하가 되기 때문이다. KoPoCo에서 안자산의 고유명사 사용 비율 순위는 전체 344명 중 16위이고, 비교 대상인 100명 시인 중에서는 단연 1위다.

안자산의 E등급 시어는 모두 165회 나오는데 그 비율(5.9438%)은 비교 대상 100인 중에서 가장 높다. D등급 시어 역시 함윤수 다음으로 비율이 높은데(34.7262%), 사용빈도(964회)는 함윤수(221회)보다 월등히 많다. 이러한 이유로 인하여 그의 시는 전체적으로 낮은 등급으로 평가된다. 아울러서 안자산의 작품 길이가 짧고, 서술적이기보다는 압축적이라는 점도 낮은 등급을 받는 데 기여하고 있다고 본다.

한용운은 평균적으로 가장 높은 등급의 작품을 쓴 것으로 밝혀졌다. 말하자면 그의 작품에서는 쉽고 보편적인 어휘 비중이 매우 높다는 말이다. 일단 그는 비교 대상 100인의 시인 중 A등급 어휘의 사용 빈도가 가장 높은 시인이다.(51.4130%) 아울러서 그의 작품 88편(어휘는 모두 8,245개)에서 대명사는 모두 874회 출현하는데, 그 비율(10.6004%)도 비교 대상 100명 시인 중에서 1위로 높다. 대명사가 많다는 것은 그의 작품이 압축적이기보다는 서술적일 가능성이 높다. 그의 대명사는 '나', '너', '그대', '당신', '그', '저', '우리', '누구', '자기' 등의 인칭대명사가 대부분이고, 그 외에 '저기', '이것', '저것', '그것', '무엇', '어디', '언제' 등의 지시대명사도 두루 쓰이고 있다. 이 중 553개가 A등급이고, 267개가 B등급, 54개가 C등급이다.

또한 그의 작품은 자유시로서는 매우 긴 편에 속한다. 어휘의 양이 많아질수록 각 등급의 어휘가 양적으로도 많아진다. 그런데 이처럼 긴 작품은 보다 서술적이어서 A등급의 어휘가 많이 쓰이고 따라서 작품의 등급도 높아지게 된다. 한용운은 비교 대상 100명 시인 중에서 자유시만 쓴 시인 중 제 9위로 작품의 길이가 길다.(평균 93.69개/작품당)[14]

---

[14] 사실 통계적 분석에는 함정이 있을 수 있다. 말하자면 서술의 양이 많을수록 문장의 기능과 관련된 허사가 많이 사용되므로 어휘의 등급은 높아질 가능성이 많다. 그러나 막사 작품 어휘의 난해성을 일반적 통계 현상만으로 설명하는 것은 매우 조심해야 한다고 본다. 단 몇 개의 어휘를 해석하지 못해서 문맥의 의미, 나아가서는 시작품의 의미를 파악하기 어려운 경우도 있기 때문이다.

한하운의 경우는 작품별 등급 조사의 결과 표준편차(0.6844)가 가장 높은 시인으로 나왔다. 그 다음 순위에 있는 김상원(0.4620), 이상(0.4273)보다 월등히 높다. 한하운의 작품 중 등급이 가장 높은 작품은 <어머니>(4.7241)이고, 가장 낮은 작품은 <개구리>(1.0000)이다. 한하운의 표준편차가 이처럼 크게 나타난 것은 우선 작품의 양이 적을 뿐 아니라 <개구리>의 평균 등급이 너무 낮기 때문이다. 그 다음으로 <양녀(洋女)>의 등급이 낮은데(3.3953), <개구리>를 예외적인 것으로 취급하여 제외한다면 표준편차는 매우 줄어들게 되고(0.3472), 한하운의 전체 어휘등급은 22위로 내려가게 된다.

4.3. 어휘 등급의 편차 사례 분석

일반적으로 작품 전체 어휘에 대한 평가 결과 점수가 낮으면, 어휘 면에서 어려운 작품일 가능성이 높고, 그 반대의 경우는 어휘 면에서 이해가 쉬운 작품일 가능성이 높다. 점수가 높으면 빈도가 높고 보편적인 어휘이기 때문이다. 평가의 결과 어휘 등급이 가장 낮은 시작품은 한하운의 <개구리>였다.

가갸거겨/　　고교구규/　　그기가//
E_가갸거겨90ng　E_고교구규90ng　E_그기가90ng : 1-1-1

라랴러려/　　로료루류/　　르리라//
E_라랴러려90ng　E_로류루류90ng　E_르리라90ng : 1-1-1

이 작품은 2연 6행으로 되어 있으며, 개구리의 울음소리를 초등학교에서 아이들이 처음 한글을 배울 때의 소리에 빗대어 표현한 작품이다. 여기 사용된 모든 어휘는 『표준사전』에 표제어에 들어 있지 않으며, 『보고서』에도 포함되어 있지 않다. 따라서 모두 E등급을 부여 받았고, 평균 등급 점수는 1.00이 되었다. 이는 매우 예외적인 현상이라 할 수 있다.

다음 작품은 김동명의 <北平>이란 작품이다.

北平은/　　헌신짝/　　二十九路軍은　　哲學者같이　賢明하다.
E_북평90nm　D_헌신짝00ng　E_이십구로군90nm　C_철학자00ng　D_현명하다01va : 1-2-1-3-2

이 작품의 평균 등급은 1.8이다.((1+ 2+ 1+ 3+ 2)/5) 특히 이 작품에는 한자 어휘가 많고(철학자, 현명하다), 고유명사가 많아서(북평, 이십구로군

) 작품 어휘 등급이 매우 낮은 것으로 나온 것이다.[15]

반면, 어휘 등급이 가장 높은 작품은 김광섭의 <獄窓에 기대어>이다.

하늘로 하늘로/ 가는 마음//
A_하늘01ng A_하늘01ng A_가다01㉠vv A_마음01ng : 5-5-5-5

맑은 바람/ 타고 가면//
A_맑다01va A_바람01㉠ng A_타다02vv A_가다01㉠vv : 5-5-5-5

흰 구름/ 눈물 씻는다//
B_희다00va A_구름01ng A_눈물01ng A_씻다00vv : 4-5-5-5

이 작품은 '희다'만 빼고 모두 A등급의 어휘로 되어 있다. 따라서 평균 등급은 4.9167(59/12)에 이른다.[16]

4.4. 시인의 어휘 등급 대비 분석- 소월과 백석의 경우

한국의 대표시인인 소월과 독특한 언어세계를 보이고 있는 백석 시인을 대상으로 어휘등급의 통계치를 대조적으로 분석해 보겠다. 대략적인 통계 결과는 다음과 같다.

표 7 작품별 어휘등급 통계 자료

분류	통계치	100 명 시인	소월	백석
분석 대상	작품수	7,494 편	145 편	36 편
	어휘수	591,077 개	7,949 개	2,219 개
작품별 어휘등급	평균값 average()	3.7580	3.9487	3.5806
	표준편차 stdev()	0.3470	0.3065	0.2831
	중간값 median()	3.7634	3.9615	3.5965
	최빈값 mode()	4.0000	4.0000	3.3333
	최대값 max()	4.9167	4.6667	4.2079
	최소값 min()	1.0000	3.2083	3.0000

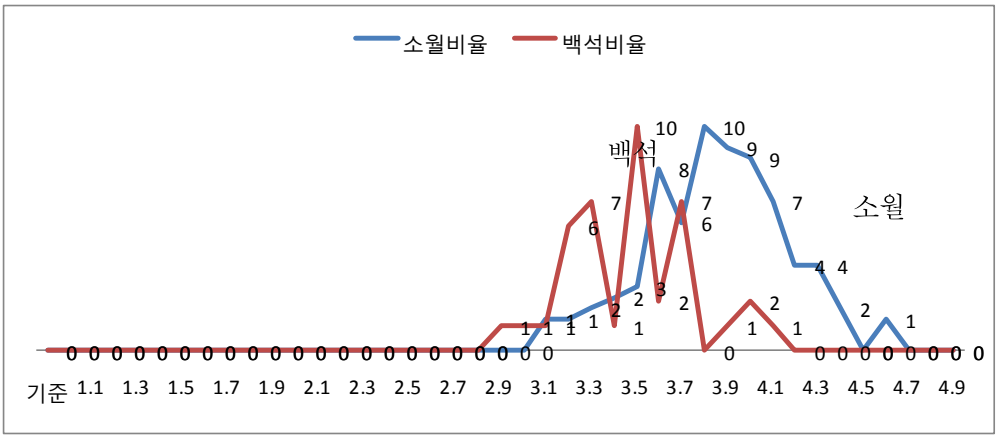
[15] 이 작품에서는 은유(북평=헌신짝)와 직유(철학자같이 현명하다)가 교차하여 사용되고 있다. '헌신짝'이란 말 자체가 비유적인 의미로 사전에 수록되어 있어서 사은유(dead metaphor)에 가까운 말이라고 할 수 있다. 향후 이러한 비유어의 참신성 혹은 매재(vehicle)와 취의(tenor)의 긴장감 등에 따라서 어휘 등급을 재산정할 필요가 있다.

[16] 제목의 '옥창00ng'은 D등급이고, '기대다01vv'는 C등급이다. 따라서 교재 편찬을 위한 작품의 어휘 등급 평가에는 제목의 어휘까지 포함하는 것이 옳다고 본다.

KoPoCo의 상위 100명 시인을 모집단으로 하는 통계치와 비교해 볼 때, 소월은 평균보다 높은 등급의 작품을 주로 썼고, 백석은 이와 반대로 평균보다 낮은 등급의 작품을 썼다고 할 수 있다. 일반적으로 말하자면 소월시는 평균보다 쉽게 이해할 수 있는 어휘로 되어 있는 반면, 백석시는 평균보다 이해하기 어려운 어휘로 되어 있는 것이다.

어휘등급의 편차 면에서는 소월이나 백석이 모집단의 평균보다 적은 편차를 보이고 있다. 비교적 일정한 등급의 어휘를 주로 구사하고 있다고 말할 수 있겠으나, 분석 대상의 양이 많으면 편차가 커지는 일반적 현상을 고려해야 한다.[17] 평균값의 경우와 마찬가지로 소월은 중간값이 모집단에 비해 높은데 비해 백석은 모집단보다 낮다. 소월과 백석을 상대적으로 비교해 볼 때에도 전체적인 통계치가 '소월 > 백석'의 현상을 보이고 있다. 이는 소월과 백석의 작품 어휘 등급의 평균 빈도를 나타내는 그래프를 통해서도 쉽게 확인할 수 있다.

그림 2 소월과 백석의 작품어휘 등급 평균 분포도



이러한 현상은 일반적으로 알고 있는 문체 인상과 그리 다르지 않다고 본다. 소월은 친근하고도 쉬운 한국어를 구사하여 한국인의 정서를 잘 나타내는 시인인데 비하여, 백석은 특수한 지역어를 풍부하게 구사하여 지역의 토착 정서를 잘 나타내는 시인이라는 것이다.

[17] 표준편차는 표본의 분량에 따라서 달라질 수 있으므로, 보다 정교한 편차 분석은 다음 연구로 미룬다.

5. 결론

지금까지 국립국어원의 등급을 적용하여 현대시의 시작품의 어휘 등급을 부여하는 방안에 대해 논의해 보았다. 국립국어원의 등급체계에 포함된 어휘는 일반적인 환경에서 사용된 어휘들이다. 따라서 시작품과 같은 장르적 특성을 반영할 수 있는 등급체계가 필요하다고 생각한다. 사실상 이 연구에서 D등급으로 분류한 어휘 중에도 고빈도어(일반 국어에서는 저빈도어인)가 있고, 또 중요 어휘들도 있기 때문에 문학 장르의 등급 분류를 위한 보다 세밀한 어휘 등급 분류 체계가 필요하다고 생각한다. 국립국어원의 경험적 등급 분류 방식을 개선하여 어휘 등급 체계에 대한 보다 객관적인 접근이 이루어져야 한다.[18] 문학의 어휘에 관심이 있는 학자들이 이 논의에 참여해 주었으면 한다.

어휘 등급 체계를 바탕으로 하는 작품의 언어적 등급을 판정하는 데 있어, 이 연구에서는 일단 현대 표준어에 의한 정상화 처리를 거쳐서 단지 어휘의 난이도에 따라 가중치를 부여하고, 그 평균값을 중심으로 검토해 보았다. 일반적으로 통계적 현상이 연구자들의 문체 직관에서 크게 벗어나지 않는 것으로 확인되었으나, 가중치를 어떤 값으로 부여할 것인지, 통계 현상을 보다 타당한 것으로 만들기 위한 다른 변수는 없는지(예를 들면 작품의 길이나 어종의 다양성 등을 변수로 적용하는 문제)에 대해서 보다 깊이 있는 연구가 필요하다고 본다. 또한 정상화의 수준에 대한 언어학적 검토라든지, 이후 통계 작업에 있어 분석의 기법을 보다 발전시키는 것도 하나의 과제가 된다.

나아가서는 작품의 어휘적 등급뿐만 아니라 표현 기교나 문학적 장치의 난이도를 고려한 작품의 종합적인 난이도에 대한 등급 체계를 개발할 필요가 있다. 앞으로 이 연구를 계속하여, 보다 정교한 어휘 등급 분류 체계를 개발하고, 이를 수필이나 소설 장르의 작품에도 적용해 볼 계획이다.

아울러서 학습자의 문화적 특성에 따른 작품 이해의 난이도를 측정하는 것도 고려할 필요가 있다. 예를 들어 같은 한자문화권인 중국과 일본, 대만 등의 학습자를 위해서 한자어로 된 시어에 대해서, 학습 효과의 전이

---

[18] 이를테면 어휘의 어원적 구성(고유어, 한자어, 외래어 등)에 대한 고려도 필요하고, 표준어, 방언, 외래어, 비속어, 슬랭 등의 어휘 성격에 대한 고려도 필요하다. KoPoCo-2011에는 표제어의 어휘 구성 분석 정보가 포함되어 있어서 향후 이를 바탕으로 하는 연구도 시도해 볼 것이다.

가 긍정적·부정적으로 일어날 가능성에 대해서 자세한 연구가 필요하다.[19]

참고문헌

· 김병선 외(2001), 『한국 현대시어 용례사전』, 누리미디어 KRPIA.
· 김병선(2004), 한국 현대시 데이터베이스의 구성과 그 활용방안, 『한국언어문학』 53집, 한국언어문학회. pp.513-535.
· 김병선(2006), 현대시인의 문체적 지문을 찾아서, 『국어국문학』 143호, 국어국문학회, pp.153-188.
· 김병선(2007), 시적 유사성 탐구 방안 연구, 『조선-한국학 국제학술대회 발표논문집』, 중국 연변대 아세아연구소.
· 김병선 외(2007), 『한국 현대시어 빈도사전』, 한국문화사.
· 김병선(2009), 시어의 기본형을 찾아서 -목록 참조를 통한 지능적 기본형 추출 방안-, 『문학 연구와 정보과학 학술회의 발표논문집』, 한국학중앙연구원 어문생활사연구소. pp.5-28.
· 김병선(2010), 문체 연구와 코퍼스의 활용, 『차세대 어문정보학의 전망 학술회의 논문집』, 한국학중앙연구원 어문생활사연구소. pp.33-55.
· 박범조(2000), 『현대 통계학 이론과 활용』, 시그마 프레스.
· 윤여탁(1999), 문학을 활용한 한국어 교육 방법, 『국어교육연구』 6집, 서울대 국어교육연구소. pp.239-256.
· 정병헌(2004), 외국인을 위한 한국문학교육의 현황과 개선 방안, 『어문연구』 44집, 어문연구학회. pp.327-350.
· 조남호 편(2002), 현대 국어 사용 빈도 조사 –한국어 학습용 어휘 선정을 위한 기초 조사, 국립국어원.
· 조남호 편(2003), 한국어 학습용 어휘 선정 결과 보고서, 국립국어원.
· 조현용(2000), 『한국어 어휘교육 연구』, 박이정.
· 최길시(2000), 『외국인을 위한 한국어 교육의 실제』, 태학사.
· 한국문학교육학회 편(2010), 『정전(正典)』(문학교육총서 2), 역락.
· 홍서연(2001), 외국어로서의 한국어 교육을 위한 문학 텍스트 분류 방안, 아주대교육대학원 석사논문.

[19] 필자가 부여한 작품 등급 정보는 자료의 양이 많아서 논문에 포함시키지 못했다. 별도의 파일로 학계에 공개하기로 한다.

· 李紹山 편(2008), 『語言研究中的 統計學*(Basic Statistics in Language Studies)*』, 西安: 西安交通大學出版社.
· Manning, Christophr D. and Sch tze, Hinrich(2000), Foundations of Statistical Natural Language Processing, The MIT Press.
· Muller, Ch.(1992), Initiation aux M thodes de la Statistique Linguistique, Paris: Hachette Universite. (배희숙 역(2000), 『통계언어학 입문』, 태학사)

Linguistic Grade To Korean Modern Poetry And Literary Education

Abstract

This study aims to level of verbal grade to Korean modern poetry by measuring the verbal difficulties of the individual words of poetry and to apply such grade system to s elect Korean modern poetic works as materials of literary text for foreign students. Th is study concentrates on the objective and systematic principles and processes for selecting such poetic texts to the foreign students of Korean literature. It adapts the computational and statistical approach to the material.

The subject of this study is the Korean modern poetry corpus of six hund red and ten thousand vocabularies in size which is gathered and managed by the author. Basically this study accepts the grading system of verbal diffic ulties of general texts in three leve ls which is developed by The National In stitute of the Korean Language, and expands it to the five grading levels, fro m A level to E level, in considering the literary texts.

To decide verbal grade of individual poetic work, this study tries to norma lize the words of diversities and gives weight score to the words according to level of difficulties and finally counts the averages, standard deviations an d other statistical factors of the score of the works. As a result such quantitat eeive phenomenon reinforces the literary intuitions of the literary scholars. Although the high frequency words of general texts are related with those of poetic works in general, some words, such as ‘im(임)’, ‘guedae(그대)’, ‘dan gsin(당신)’, and so on, are confirmed as the poetic words by the discrete gra de score with the words of general texts.

Above all the 100 poets, Han Yongun(韓龍雲) writes the most highest ver bal level of poetic works, so to speak the most easiest works. In most poems in his anthology ‘The Silence of the Beloved’, he prefers some discursive se ntences. On the other, An Jasan(安自山) writes the most lowest verbal level of poetics works, and his poems is only written with sijo form. So the decisi ve factors that related with the verbal difficulties of poetry are genres (free p oem, prose poem. sijo, dramatic and epic poem) and the length of the poems . In contrasting Kim Sowol(金素月) and Baekseok(白石), the former acquir es the higher score than the latter. In other words, Sowol who takes normal and general words as his poetic words to present traditional emotion of Ko

rean writes easier poems than Baekseok who indulges in presenting the rural emotions of his own.

It is needed that more sophisticated statistical analysis to the verbal level of literary vocabularies should be considered. And also the linguistic approa ches to the normalization of verbal variables should be developed. (c) kimbs @aks.ac.kr

Keywords：Korean Modern Poetry, Literary Education, verbal grade