Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 4, pp. 1057 ~ 1064

Quantitative Linguistic Analysis on Literary Works

Kyung-Ho Choi[1)]

Abstract

From the view of natural language process, quantitative linguistic analysis is a linguistic study relying on statistical methods, and is a mathematical linguistics in an attempt to discover various linguistic characters by interpreting linguistic facts quantitatively through statistical methods.

In this study, I would like to introduce a quantitative linguistic analysis method utilizing a computer and statistical methods on literary works. I also try to introduce a use of SynKDP, a synthesized Korean data process, and show the relations between distribution of linguistic unit elements which are used by the hero in a novel 'Sassinamjunggi'and theme analysis on literary works.

Keywords : 계량언어학, 문학작품, 자연언어처리, 통계적 방법

1. 서 론

작가들은 저마다 고유한 문체로 글을 쓴다. 그리고 각 작품들은 고유한 테마를 지니다. 그런데 이러한 문체나 테마는 대단히 추상적이어서 분명한 실체를 찾아내기 어렵다. 많은 문학 연구가들은 작품 속에 숨겨져 있는 코드를 여러 가지 색채로 풀어낸다. 이들의 분석을 통해 같은 작품이 품고 있는 다양한 모습들이 하나씩 드러난다(배희숙, 1999). 이 때 연구대상인 문학작품의 특성이 연구자의 주관적인 견해나 관점을 통하여 해석됨과 동시에, 좀 더 객관적인 방법으로 나아가 데이터까지 함께 제시되는 연구가 수행된다면 보다 바람직하다고 할 수 있겠다.

기존 언어학 연구에서의 일반적인 연구방법은, 모국어 화자의 직관에 의존한 논리적 설명이 대부분이었다. 그러나 최근 연구는 직관에 의존한 연구에서 대규모의 실제적 자료를 이용하는 계량적 연구로 옮아가고 있다(최경호 · 황용주, 2007). 사실 언어학에 대한 계량적 연구는 오래 전부터 시도되어 왔다. 특히 최근에는 컴퓨터의 발달

1) Professor, Department of Data Science, Jeonju University, Jeonju, Jeonbuk, 560-759, Korea
E-mail : ckh414@jj.ac.kr

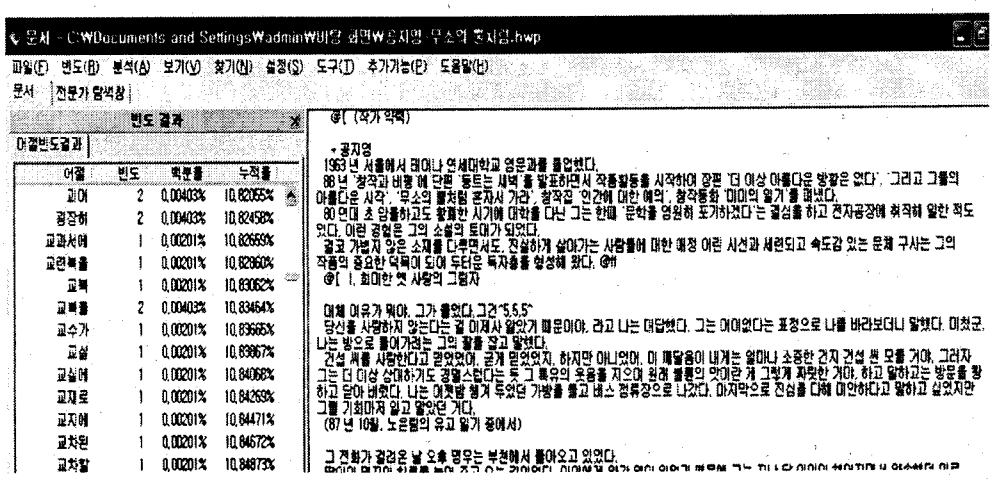
과 함께 인간의 언어체계를 컴퓨터에 인식시켜 인간과 인간, 인간과 기계 사이의 자유로운 의사소통을 위한 연구가 활발해지고 있다. 이와 더불어 인간이 직관적인 방법으로 포착할 수 없는 미세한 언어 단위들의 특성이나 단위들 간의 관계를 통해 언어의 내적 구조를 연구하는 계량언어학에 대한 관심도 높아지고 있다(배희숙, 2000).

계량언어학이란 자연언어처리 관점에서 보면, 통계적 방법에 의존하여 언어를 연구하는 언어학의 한 분야로서, 언어적 사실을 주로 통계적 방법에 의하여 양적으로 해석함으로써 언어가 지니는 여러 성질을 밝혀내려고 하는 계산언어학의 한 분야이다. 나아가 계량언어학이란 국어정보학의 관점에서 보면, 코퍼스(말뭉치, corpus)를 구성하고 계량화한 뒤 유의미한 계량단위에 대한 측정의 결과를 통계학적으로 분석하여 코퍼스에 담긴 내용의 성격과 코퍼스 자체의 성격을 비롯한 각종 의미를 규명하는 언어학의 한 분야이다(임칠성, 2003).

이렇듯 언어연구에서 컴퓨터와 통계적인 방법을 활용함으로써 수작업에서 생길 수 있는 오류와 개인의 주관적인 판단을 최소화하고, 과학적 · 객관적인 방법으로 연구할 수 있는 이점을 갖게 된다. 이에 본 연구에서는 문학작품에 대한 컴퓨터와 통계적인 방법을 활용하는 계량언어학적 분석방법을 소개하고자 한다. 이 과정에서 통합형 한글 데이터 처리기 SynKDP(소강춘, 2002)의 활용을 소개하고, 소설 '사씨남정기'를 대상으로 주인공이 사용하고 있는 언어 단위 요소의 분포가 문학작품의 테마분석과 어떤 관계를 맺고 있는지를 보여주고자 한다.

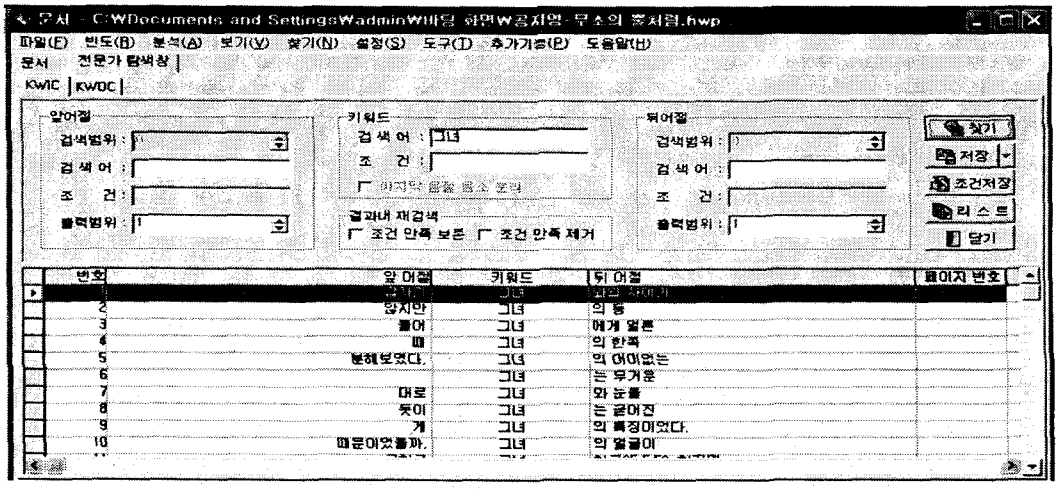
2. 정보처리 프로그램 SynKDP 소개

일명 깜짝새라고도 불리는 SynKDP는 소강춘(2002)에 의하여 개발된 통합형 한글 데이터 처리기로, 구축된 말뭉치를 이용하여 정보화에 초보적인 연구자라도 원하는 목적에 따라 언어학적 정보를 신속하고 정확하고 편리하게 얻어낼 수 있는 도구이다. 이 프로그램의 화면구성은 메뉴창, 문서창, 그리고 결과창 등으로 구성되어 있는데, 주요기능이 모여져 있는 문서창은 <그림 1>과 같다.



<그림 1> SynKDP 문서창의 화면구성

<그림 1>은 공지영의 무소의 뿔처럼에 대한 어절빈도분석 결과를 보여주고 있다. 문서창은 자료를 불러올 뿐만 아니라 열려진 파일에 대한 모든 검색이 버튼 클릭만으로 처리될 수 있다. 문서창에서 '전문가 탐색창'을 선택하면 <그림 2>와 같은데, 이 부분이 SynKDP의 가장 핵심적인 부분이라 할 수 있다. 여기에서는 키워드, 앞어절, 뒤어절 등에 대한 검색조건 및 출력조건을 입력함으로써, 현대한글, 옛한글, 이두, 구결, 한자, 영문 등 국어학 연구에 필요한 모든 문자검색이 가능하다.



<그림 2> KWIC 검색결과

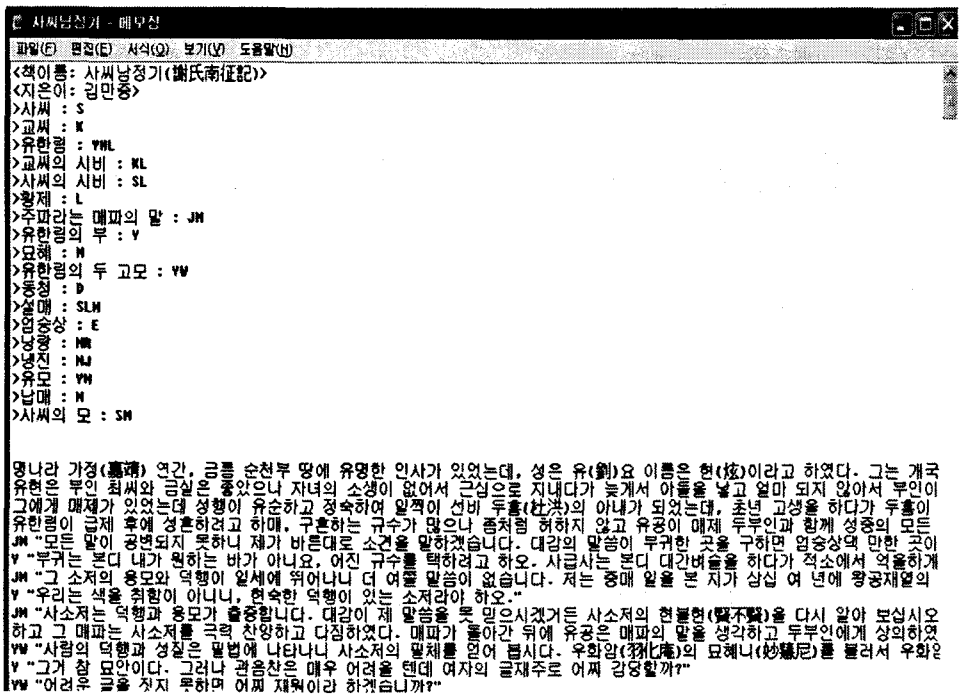
3. 계량언어학적 분석 예

언어학은 언어의 특징을 밝혀냄을 그 연구 목적으로 하지만, 언어정보학은 언어에 담겨져 있는 모든 내용을 정보화하여 이 중에서 연구자들이 요구하는 모든 정보를 추출해 낼 수 있도록 하는 것을 연구목적으로 한다. 이러한 언어정보학의 한 분야인 계량언어학에서는 컴퓨터의 도움을 받아 통계적인 방법을 사용하여 계량화된 분석을 실시한다는 점이 특징이라 할 수 있다. 한 예로, 김소월의 시에 나타나는 언어의 소리가 어떤 특성이 있는지에 관심이 있다고 하자. 즉 김소월의 시의 언어가 서정성을 갖는다면, 그의 시에 쓰이는 소리가 그 서정성을 반영하는 어떤 특성이 있는지가 궁금할 수 있다(강범모, 2003). 이 때 코퍼스를 구축하고 상기 언급된 SynKDP 등과 같은 문서처리기를 활용한다면, 눈 깜짝할 사이에 원하는 결과를 얻을 수 있다. 이러한 점이 그동안 언어학 연구와 다른 계량언어학의 특징이라 하겠다.

본 연구는 문학작품에 대한 계량언어학적 분석을 소개하고자 하는데 그 일차적인 목적이 있다. 이에 본 연구에서는 김만중의 사씨남정기(김명환, 1997)를 분석대상 작품으로 선정하고 계량분석을 실시해 보았다. 문학작품에 대한 계량분석을 위한 일반적인 절차는 크게 양화작업과 통계작업으로 구분되는데, 양화작업은 다시 텍스트 입

력(1단계), 분할작업(2단계), 어휘정리작업(3단계) 등으로 구분된다(배희숙, 1999).

이상의 과정을 거쳐 <그림 3>과 같이 구축된 코퍼스에 대한 계량분석을 한 결과가 다음과 같다. 먼저 <표 1>은 두 주인공인 '교씨'와 '사씨'의 언어에 나타난 어절빈도수 결과이다.



<그림 3> 구축된 코퍼스(일부)

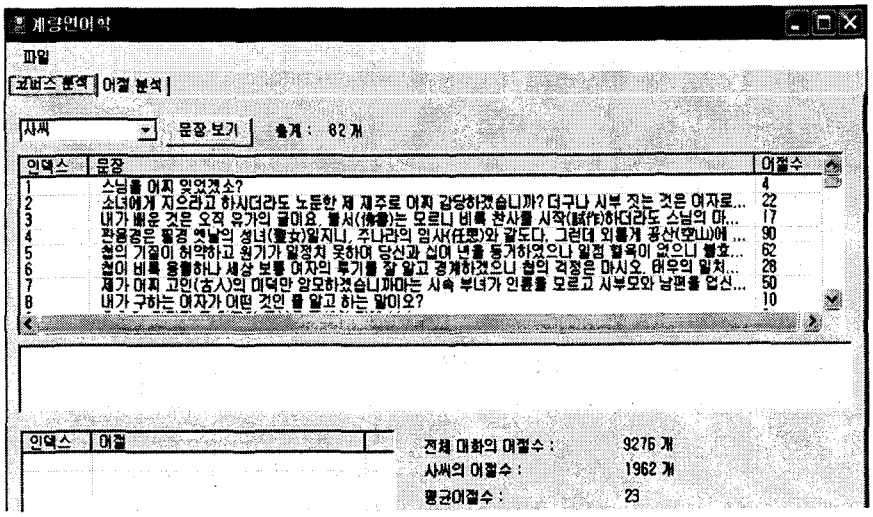
<표 1> '교씨'와 '사씨'의 언어에 나타난 어절빈도수

순위	교씨		싸씨	
	어절	빈도수	어절	빈도수
1	나	32	나	68
2	하다	24	어떻게	35
3	유한림	19	하다	30
4	사씨	17	있다	27
5	어떻게	17	유한림	26
6	일	17	것	24
7	않다	14	못하다	24
8	좋다	13	사람	23
9	있다	12	없다	19
10	것	11	가다	18
11	그	10	몸	17
12	없다	10	않다	17
13	이번	10	일	16
14	되다	9	말씀	15
15	방법	9	죄	14
16	아기	9	되다	13
17	계교	8	묘혜	13
18	말씀	8	시부	12
19	사람	8	죽다	12
20	생각	8	그러나	11
21	주다	8	구하다	10
22	두부인	7	그	10
23	말다	7	말다	10
24	못하다	7	아니다	10
25	버리다	7	여자	10

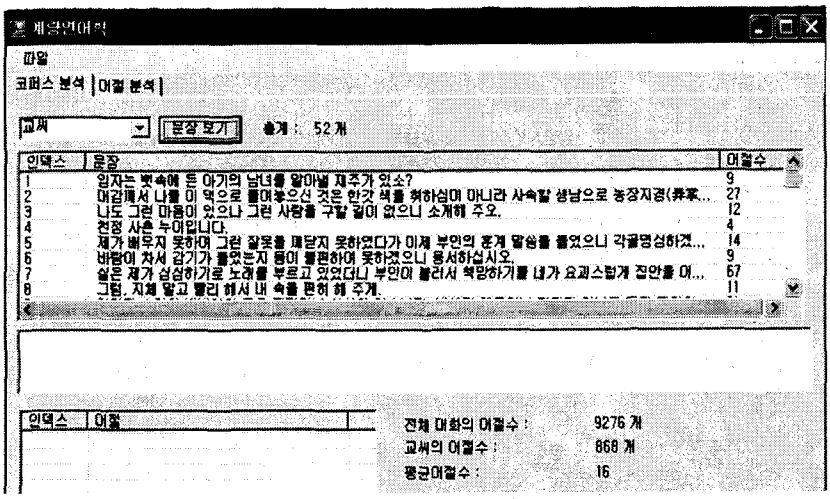
여러 가지 관점에서 해석해 볼 수 있겠으나, 일단 <표 1>로부터 '사씨'나 '교씨' 모두 자기 자신을 지칭하는 어휘를 많이 사용하고 있다. 특이한 점은 '교씨'에 비하여 '사씨'에게서 '어떻게'라는 어휘가 상대적으로 높게 나타나는데, 이는 작품의 테마와 관련된 때문으로 해석된다. 사씨남정기의 역사적 배경을 보면 널리 알려진 대로, 숙종(유한림)이 인현왕후(사씨)를 폐위하고 장희빈(교씨)을 왕비로 맞아들이는데 반대하여 유배지에서 쓴 작품이다. 작품 속에서 마음씨 고운 '사씨'는 남편에게 순응하고 남을 음해하거나 모략함이 없이 자기 자신의 신세를 한탄하는 모습으로 그려져 있다. 이에 '사씨'에게서 '어떻게'라는 어휘빈도가 높게 나타나는 것은, 작품의 테마와 아주 밀접한 결과라 할 수 있다.

전체 어절 수가 9,236개인 가운데, '사씨'가 사용한 어절 수는 1,962개이며, '교씨'가 사용한 어절 수는 '868'개로 나타났다. 이는 '교씨'와 '사씨'의 출신성분에 기인한 것으로, '교씨'보다 '사씨'가 보다 폭 넓은 어휘를 활용한 결과라 해석되는 대목인 바, 역시 작품의 테마와 관련된 것으로 해석할 수 있다. 한편 <그림 4>와 <그림 5>는 '사씨'와

'교씨'의 평균어절 수를 나타낸 것이다. 각 대화에서 사용한 '교씨'의 평균 어절 수가 16개인 반면 '사씨'가 사용한 평균 어절 수는 23개로, '사씨'가 훨씬 장문의 대화를 하고 있음을 알 수 있다.



<그림 4> '사씨'의 코퍼스분석



<그림 5> '교씨'의 코퍼스분석

이제 작품 속에서 '사씨'와 '교씨'의 대화를 두 독립집단으로 가정하고, 집단에 따라 대화중에 나타난 본인과 상대 그리고 유한림을 지칭(호칭)한 비율이 통계적으로 유의한 차이가 있는지 (동질성)검정해 보자. 먼저 실측빈도수는 <표 2>와 같다.

<표 2> '사씨'와 '교씨'의 대화에서 주인공들을 지칭(호칭)한 빈도수

화자 \ 피 지칭인	본인	상대	유한림	합계
사씨	68	7	26	101
교씨	32	17	19	68
합계	100	24	45	169

검정결과 검정통계량 χ^2=59.57(p=0.000)로 유의수준 5%에서 집단에 따라 대화중에 나타난 본인과 상대 그리고 유한림을 지칭(호칭)한 비율이 통계적으로 유의한 차이가 없다는 귀무가설이 기각되는 것으로 나타났다. 이에 '사씨'와 '교씨'의 대화에서 주인공들을 지칭한 비율이 동일하지 않음을 알 수 있다.

4. 결론

계량언어학은 분석의 과정에서 통계적 방법을 활용하기에, 통계언어학으로 지칭되기도 한다. 통계언어학은 코퍼스를 구성하는 언어 단위들의 관계를 통해 언어의 내적 구조를 파악하는 학문으로, 순수 언어 연구뿐만 아니라 그 응용분야인 언어정보처리, 자연언어처리 등에 많은 도움을 주고 있다(배희숙, 2002).

본 연구에서는 통계적인 방법을 활용하여 문학작품에 대한 계량언어학적 분석의 과정을 보여주었으며, 그 과정에서 문학작품에 대한 해석의 또 다른 양상을 제시하였다. 물론 계량언어학적 분석이 기존 언어학에서의 연구를 압도하는 전혀 새로운 것은 아니다. 그러나 계량언어학적 분석에서 도출된 결과를 언어학 연구의 기초자료로 활용한다면, 보다 풍부하고 객관적으로 문학작품에 대한 이해와 해석이 가능해질 것으로 사료된다. 이에 기존의 언어학 연구에 통계적인 방법을 토대로 하는 계량언어학 연구가 상호보완적으로 활용될 수 있기를 바라는 바이다.

참고문헌

1. 강범모 (2003). 언어 컴퓨터 코퍼스 언어학, 고려대학교출판부.
2. 김명환 (1997). 한국대표고전소설3, 빛샘.
3. 배희숙 (1999). 문학작품의 양적 분석과 컴퓨터의 활용, 학술발표와 연구자료집, 2, 프랑스문화예술학회.
4. 배희숙 (2000). 어휘 풍부성 평가에 대한 계량언어학적 연구, 음성과학, 7(3), 139-149.
5. 배희숙 (2002). 통계언어학의 원칙과 방법, 한국어와 정보화, 태학사.
6. 소강춘 (2002). 정보처리 프로그램에 대하여, 한국어와 정보화, 태학사.
7. 임칠성 (2003). 5·18항쟁 관련 유인물과 성명서 어휘의 계량연구(1), 계량언어학 2집, 박이정.

8. 최경호, 황용주 (2007). 계량언어학 연구에서 통계적방법의 활용. 한국데이터정보과학회지, 18(2), 269-278.

[2007년 11월 접수, 2007년 11월 채택]