Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 2, pp. 269 ~ 278

Application of Statistical Methods in Quantitative Linguistics Study

Kyung-Ho Choi[1] · Yong-Joo Hwang[2]

Abstract

Nowadays, from the study of quantitative linguistics, the application of quantitative method is located in a variety of fields as a necessary method. According to this phenomenon, the knowledge of statistical method is requisite for linguists. However, unfortunately, there still remain difficulties for them to acquire the statistical knowledge. So, it is needed for linguists to be helped by statisticians and their active roles. Accordingly, this study is going to emphasizing that statisticians should have more interests in the field of quantitative linguistics. Moreover, it will prove that by using statistical methods, analysis on the linguistic research becomes more objective and scientific.

Keywords : 계량언어학, 국어정보학학, 말뭉치, 통계적 방법

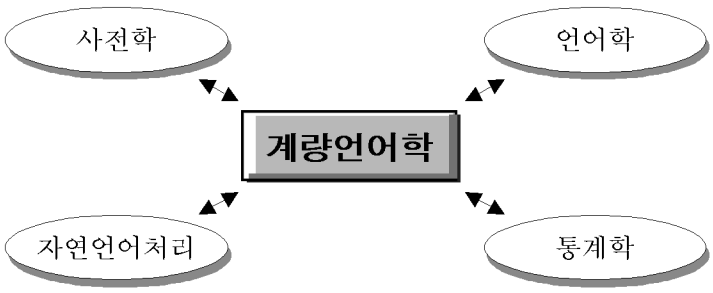
1. 언어연구에 대한 계량적 접근의 필요성

기존 언어학 연구에서의 일반적인 연구 방법은, 모국어 화자의 직관에 의존한 논리적 설명이 대부분이었다. 그러나 최근 연구는 직관에 의존한 연구에서 대규모의 실제적 자료를 이용하는 연구로 옮아가고 있다. 그런데 대규모 언어자원을 효율적으로 활용하고, 나아가 정보화 시대에 필수적으로 요구되는 언어자원의 기계적 자동처리 등을 위해서는 객관적이고 실증적인 자료를 바탕으로 언어자원을 활용할 수 있는 방법론이 필요하다. 즉 한국어 어휘와 구문의 특징을 직관이나 이론적 방법을 통해 설명하는 것과 다르게, 명시적이고 객관적으로 실제 언어자료를 활용하여 한국어의 특징을 밝히는 효과적인 연구방법이 절실히 요구된다(박병선, 2005). 그러나 현재까지 대규모 자료처리에 필수적인 통계적인 처리방법론을 언어연구에 복합적으로 활용한 계량언어학(quantitative linguistics)적 연구 사례는 그리 많지 않다.

1) Professor, Department of Data Science, Jeonju University, Jeonju, Jeonbuk, 560-759, Korea E-mail : ckh414@jj.ac.kr
2) Instructor, Department of Korean Language and Literature, Chonbuk National University, Jeonju, Jeonbuk, 561-756, Korea, E-mail : haeri@unitel.co.kr

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 통계적인 방법을 활용함으로써 수작업에서 생길 수 있는 오류와 개인의 주관적 판단을 최소화하고, 과학적 · 객관적인 방법으로 담론을 분석할 수 있는 이점을 갖게 된다.

국어정보학의 연구방법을 제시한 서상규와 한영균(1999)에서는 기존의 언어학과는 달리 철저히 자료를 기반으로 하는 연구영역으로, 코퍼스 안에서의 각 언어 단위들의 빈도(frequency)와 분포(distribution) 그리고 연어 관계(collocation relation) 등을 밝히는 일이, 의미나 기능을 밝히는 일 못지않게 중요한 과제가 되며, 결국 이들 단위의 통계적 특성을 밝히기 위한 방법론의 개발이 또 다른 중요한 과제의 하나가 된다고 하였다. 그들에 따르면 계량언어학 연구에서 무엇보다도 중요한 세 가지 축은, '코퍼스'와 이를 가공처리하기 위한 '컴퓨터' 그리고 추출된 언어정보의 '통계분석'이라 할 수 있다. 그래서 그들은 계량언어학과 인접학문과의 연계관계를 <그림 1>과 같이 나타내고 있다.



<그림 1> 계량언어학과 인접 학문과의 연계관계

이상을 통해 볼 때, 대규모 언어자료를 다루는 계량언어학에서는 전산처리와 더불어 통계적 방법론의 활용 및 한국어에 맞는 통계적 방법론의 개발이 필수적인데, 아직 이에 대한 연구가 많이 이루어지지 않은 실정이다. 현재까지 이루어진 대표적인 연구사례는 박병선(2005), 남윤진(2000), 이상억(2003), 홍종선(2003), 강범모(1999), McEnery, T., Wilson, A. (1998), Kilgariff, A. (1996) 등을 들 수 있다. 그런데 국내 연구의 경우 현재까지 계량언어학을 이끌어온 중심 연구자들은 거의 언어학자들이다. 계량언어학 연구에 기여한 통계관련 학자로는 허명회(1998)와 양경숙(2003) 그리고 황용주와 최경호(2006) 등에 불과하다.

오늘날 통계적 방법의 적용은 계량언어학의 다양한 분야에 걸쳐 필수적 방법으로 자리 잡아가고 있다. 주로 대규모 자료처리를 기본으로 하는 계량언어학의 특성 상 수작업으로 자료를 처리한다는 것은 불가능하고, 또 그 처리한 결과를 효과적으로 이용하기 위해서는 통계 기법의 적용이 필수적이다. 이에 언어학자들에게 있어 통계학적 지식은 반드시 요구되는 사항이 되었다. 그러나 짧은 시간 내에 언어학자가 통계적인 지식을 모두 습득하기에는 어려움이 많다. 이에 통계학자의 도움과 함께 적극적

인 역할이 요구된다. 이에 본 연구에서는 첫째, 계량언어학 연구 분야에 통계관련 학자들이 더욱 적극적인 관심을 보일 필요성을 강조하고 둘째, 통계적 방법을 활용함으로써 언어연구의 해석을 보다 객관적이고 과학적으로 할 수 있음을 보이고자 한다.

2. 언어연구에서 활용되고 있는 통계적인 방법들

지금까지 계량언어학 연구에서 활용되고 있는 대부분의 통계 방법들은 기술통계적인 수준을 크게 벗어나지 못하고 있다. 일부 연구에서 다변량분석 등을 활용하고 있으나, 가장 많이 활용되고 있는 방법은 빈도목록을 작성하는 것으로 이를 통하여 해당 텍스트의 성격 및 주요 특징을 연구한다.

한편 계량언어학 연구에서 통계적인 방법을 가장 많이 활용하는 분야는 공기관계(co-occurrence)분야인데, 한국어의 공기관계를 중심으로 한국어 연구를 위한 다양한 계량적 접근방법과 계량언어학적 분석 방법을 제시한 연구로 박병선(2005)을 들 수 있다. 공기관계 연구에 계량적 접근방법이 필요한 예를 들어보면 다음과 같다. 즉, 연구대상이 되는 텍스트에 대해, '찬 바람'과 '차가운 바람' 중 어느 것이 더 많이 사용되는지를 알고자 하는 경우, 이에 대한 의문은 직관만으로는 정확히 알기 어렵다. 이러한 경우 사람들이 사용한 실제 자료를 이용하여, 통계적 방법을 통해 문제되는 형식의 분포와 그 유의미성을 조사하고 분석한다면 보다 객관적인 결론을 유도할 수 있을 것이다. 이를 위하여 계량언어학에서 주로 활용되는 통계량으로는 z-점수와 t-점수 등을 들 수 있다.

$$z = \frac{O - E}{\sigma} \tag{2.1}$$

$$t = \frac{O - E}{\sqrt{O}} \tag{2.2}$$

단, O : 스팬 내에서 관찰된 단어의 출현빈도
E : 같은 단어의 예상 출현빈도
σ : 전체 텍스트에서 같은 단어의 출현 표준편차

Barnbrook(1996)는 연어(collocation)에 대한 공기관계 연구에서 z-점수에 대해서는 3보다 크면 '유의적'으로, 그리고 t-점수에 대해서는 2보다 크면 '유의적'으로 판단할 것을 권하고 있다.

한편 미즈노 슌페이(2003)는 宮澤達夫(1970)의 문서 유사성 측도를 활용하여 계량적 연구를 수행하였는데, 비교하고자 하는 두 문서 A와 B에 대하여 특정어휘의 출현비율이 다음과 같을 때 유사성 측도는 식 (2.3)과 같다.

어 휘	문서 A	문서 B	차 이
w1	$p_1(A)$	$p_1(B)$	$\|p_1(A)-p_1(B)\|$
w2	$p_2(A)$	$p_2(B)$	$\|p_2(A)-p_2(B)\|$
w3	$p_3(A)$	$p_3(B)$	$\|p_3(A)-p_3(B)\|$
.	.	.	.
.	.	.	.
.	.	.	.

$$\text{유사성}=\sum_i \min[p_i(A),\ p_i(B)] \tag{2.3}$$

3. 언어연구 해석의 객관화를 위한 통계적 방법의 활용

전혜영(2005)은 일상 언어에서 '여자'와 '남자'라는 단어가 어떤 동사와 결합하여 연어를 구성하는지 알아보고자, KAIST 말뭉치 자료의 한국어 용례색인을 활용하여 자료를 수집하고, 은유의 유형과 빈도 면에서 남녀차이가 있음을 보였다. 다음 <표 1>은 그 결과의 일부이다.

전혜영은 <표 1>을 토대로, 『남녀 공통으로 들어가 있는 동사는 '얻다, 다루다, 버리다, 갖다, 고르다, 두다'로 모두 '물건'은유에 속한다. 그러나 고빈도 은유동사에서 남녀차이가 분명하게 나타나고 있다』라고 주장하고 있다. 그런데 이러한 주장이 좀 더 객관적이려면, 개인적인 판단보다는 계량화된 측도를 활용하는 것이 더욱 합리적이라 판단된다.

<표 1> 남녀 대비 은유 구성 동사 중 고빈도 동사

	여자의 경우	은유동사 424개 대비 상대비율	남자의 경우	은유동사 116개 대비 상대비율
동사(빈도)	얻다(28)	0.066	고르다(9)	0.077
	사다(19)	0.045	얻다(9)	0.077
	다루다(17)	0.040	택하다(6)	0.052
	버리다(17)	0.040	두다(5)	0.043
	갖다(15)	0.035	선택하다(5)	0.043
	고르다(14)	0.033	잡다(5)	0.043
	두다(13)	0.031	소유하다(4)	0.034
	잡다(13)	0.031	갖다(3)	0.026
	끼다(13)	0.031	다루다(3)	0.026
	끌다(11)	0.026	버리다(3)	0.026
	구하다(10)	0.023	정복하다(3)	0.026
	풀어주다(10)	0.023	내놓다(3)	0.026
	차지하다(9)	0.021	던지다(3)	0.026
	팔다(9)	0.021	사로잡다(3)	0.026

그래서 <표 1>의 자료에 대해 식 (2.3)의 유사성 측도를 활용하여 유사도를 측정하여 보면 0.24로 비교적 작은 값이다. 따라서 이를 토대로 은유 동사에서 남녀 차이가 크지 않다고 주장하는 것이 객관적 측면에서 훨씬 설득력이 높다.

또 다른 예로 황용주(2006)를 고려해 보자. 황용주(2006)에서는 신소설(1910년 대 전후) 자료에서 연어 구성이 어떠한 양상으로 나타나고 있는지 알아보고, 나아가 현대국어 말뭉치(1990년대 이후)와 비교하여 연어 구성의 변화 양상을 살펴보았다. 이 과정에서 위에서 언급된 황용주는 식 (2.1)과 (2.2)를 사용 하는 등 비교적 계량화된 방법을 활용하여 설명하고 있다. 황용주(2006)의 논의에서 구체적인 통계방법을 활용하고 있지는 못한 단점을 황용주(2007)에서는 극복하고 있다. 다음 <표 2>는 황용주(2007)에서 제시한 '담배'와 관련된 연어구성 자료이다.

<표 2> 신소설과 현대소설에 나타난 '담배' 관련 연어빈도

연 어	신소설 빈도	현대소설 빈도
먹다	2,400	1334
피우다	28	341
태우다	6	102
빨다	20	75

<표 2>의 자료에 대해 각 셀에 대한 기대도수를 구해보면<표 3>과 같으며, 이를 활용한 분할표 검정결과 p-값이 0.000으로 유의수준 5%에서 귀무가설(H_0 : 신소설과 현대소설의 연어빈도의 발현비율이 같다)이 기각되어 신소설과 현대소설에서 연어의 발현비율이 같지 않음을 알 수 있다. 나아가 <표 3>으로부터, 신소설에서는 '먹다'의 발현빈도가 기대빈도보다 높은 반면 현대소설에서는 낮음을 알 수 있다. 이렇듯 통계적인 방법을 활용함으로써 보다 객관적이고, 내용면에서도 풍부한 해석을 할 수 있다.

<표 3> 연어빈도에 대한 기대도수

			시대구분		
			신소설	현대소설	전체
연어 구분	먹다	빈도	2400	1334	3734
		기대빈도	2128.0	1606.0	3734.0
	피우다	빈도	28	341	369
		기대빈도	210.3	158.7	369.0
	태우다	빈도	6	102	108
		기대빈도	61.5	46.5	108.0
	빨다	빈도	20	75	95
		기대빈도	54.1	40.9	95.0
전체		빈도	2454	1852	4306
		기대빈도	2454.0	1852.0	4306.0

다음으로 임칠성(2003)의 연구를 살펴보자. 그는 5·18 당시 발표된 유인물, 성명서 및 언론을 비롯한 외부의 자료 등에 사용된 어휘를 1980년 1년 동안을 시기별로 항쟁전, 항쟁 중, 항쟁 후로 나누고 계량하고 그 성격을 고찰하고자 하였다. 그리하여 어휘 계량이라는 과학적인 방법으로 항쟁의 진행 시기별로 성명서와 유인물의 성격을 규명하고자 하였는데, 내용면에서 보면 담론적인 형식을 취하고 있다. 즉 결론을 도출함에 있어 구체적인 통계적 방법을 활용하지 못한 아쉬움이 있다. <표 4>는 지역감

정과 관련된 어휘의 빈도수로, 임칠성 연구의 일부이다.

<표 4> 항쟁 시기별 지역감정과 관련된 어휘의 빈도 수

어휘	항쟁 전	항쟁 중	항쟁 후
지역감정	0	5	8
전라도	0	7	3
호남	0	1	4
지역	8	30	11
경상도	0	3	2

<표 4>의 자료에 대해 임칠성은 “항쟁 전에는 지역감정과 관련된 용어가 거의 등장하지 않다가, 항쟁 중과 항잰 후에 지역감정을 자극하는 표현들이 등장한다”라고 언급하고 있다. 시기별 지역감정과 관련된 어휘의 발현 빈도수를 보면 각각 8회, 46회, 28회로 나타나 이와 같은 주장에 문제가 없어 보이나, 주관적인 판단임에는 분명하다. 따라서 이러한 경우에도 <표 4>의 자료를 토대로 귀무가설(H_0 : 항쟁시기별 사용어휘의 발현비율이 같다)에 대한 분할표 검정을 실시(χ^2=15.857, p-값=0.044)하고 그 결과를 활용한다면, 보다 객관적인 결론을 유도할 수 있을 것으로 여겨진다.

통계적 방법을 활용함으로써 보다 더 유용한 결론을 내릴 수 있는 네 번째 예로 이정란(2005)의 연구를 고려해 보자. 이 연구에서는 한국어를 외국어로 학습하는 과정에서, ‘-어서’와 ‘-니까’의 변이를 일본어권, 중국어권, 영어권 학습자 별로 차이가 있는지를 알아보고자 하였다. 이를 위하여 수집된 자료에 대해 일부 계량적인 분석을 실시하고 있으나, 통계적인 방법을 추가적으로 활용한다면 보다 의미 있는 정보를 유도할 수 있을 것으로 여겨진다. 예컨대 각 언어권 학습자별 인터뷰 발화에 대한 정답률 자료인 <표 5>를 보자.

이 자료를 토대로 각 언어권별로 ‘-어서’와 ‘-니까’의 정답률 간의 상관관계를 분석해 보면 다음과 같다. 먼저 일본어권 학습자에 있어서는 0.373으로 가장 높고, 영어권 학습자에 있어서는 0.282 그리고 중국어권 학습자에 있어서는 0.212로 가장 낮게 나타났다. 나아가 일본어권 학습자에 대한 상관관계는 유의수준 10%에서 통계적으로 유의(p-값=0.10)하게 나타났다. 그런데 이정란의 연구에서는 이러한 정보를 바탕으로 해석을 하고 있지는 못하다. 따라서 이 연구에서도 상관분석과 같은 통계분석을 활용함으로써 보다 깊이 있고 풍부한 결론을 유도할 수 있을 것으로 판단된다.

<표 5> 언어권별 인터뷰 발화 정답률

일본어	어서 정답률	니까 정답률	중국어	어서 정답률	니까 정답률	영어	어서 정답률	니까 정답률
1	75.0	85.7	1	75.0	75.0	1	40.0	0.00
2	100	100	2	100	0.00	2	100	0.00
3	60.0	0.00	3	66.6	66.6	3	85.7	100
4	100	100	4	0.00	50.0	4	50.0	33.3
5	60.0	69.2	5	50.0	0.00	5	100	0.00
6	25.0	75.0	6	100	0.00	6	40.0	66.6
7	85.7	85.7	7	16.6	40.0	7	80.0	60.0
8	50.0	86.6	8	66.6	-	8	83.3	0.00
9	0.00	72.7	9	100	0.00	9	54.5	44.4
10	40.0	57.1	10	66.6	50.0	10	16.6	42.8
11	50.0	57.1	11	0.00	0.00	11	0.00	0.00
12	66.6	77.7	12	0.00	0.00	12	50.0	33.3
13	100	88.8	13	50.0	75.0	13	100	85.7
14	90.9	83.3	14	60.0	0.00	14	50.0	71.4
15	75.0	75.0	15	57.1	25.0	15	100	100
16	88.8	85.7	16	57.1	57.1	16	0.00	50.0
17	85.7	66.6	17	60.0	60.0	17	42.8	42.8
18	16.6	60.0	18	100	100	18	75.0	20.0
19	77.7	87.5	19	66.6	66.6	19	33.3	16.6
20	33.3	85.7	20	0.00	0.00	20	0.00	0.00

마지막으로 다양한 관점에서 외래어의 유형을 분류하고 실제 국어자료에서의 사용 양상을 밝히는 것을 목적으로, 국립국어원에서 구축한 약 100만 어절의 소설 어휘 분석 말뭉치를 대상으로 한 김한샘(2005)의 연구를 보자. 이 연구에서 김한샘은 외래어의 전문 영역별 종수와 빈도에 대한 연구를 위해 <표 6>을 작성하였다. 그리고 이를 토대로 "국립국어원의 『표준국어대사전』에서 분류한 51개의 전문 영역 중에 44개 전문 영역에 해당하는 외래어가 말뭉치에 출현했다. 그 중 가장 많은 비율을 차지하는 … … 교통, 의학, 군사 등의 순으로 외래어가 많이 쓰인다"와 같이 <표 6>에 대하여 설명하고 있는 바, 수집된 정보를 충분히 활용하고 있지 못하다. 예컨대, 각 전문 영역별 외래어에 대해 '종수 순위'와 '사용률 순위'를 보면 '건설'의 경우는 각각 10위와 1위, '교통'의 경우에는 각각 27위 8위로, 일부 전문 영역에 있어 상당한 차이를 보이고 있다. 그러함에도 불구하고 '종수 순위'와 '사용률 순위'에 대하여 스피어만의 순위상관계수를 계산해 보면, 0.881로 유의수준 1%에서 통계적으로 유의하게 된다. 따라서 이 경우에도, 이와 같이 통계적인 방법을 활용한 결과를 활용함으로서, <표 6>에 대하여 보다 유용한 결론을 유도할 수 있게 된다.

<표 6> 외래어의 전문 영역별 종수와 빈도

전문영역	종수 순위	사용률 순위	전문영역	종수 순위	사용률 순위
건설	10	1	언어	14	23
운동/오락	1	2	법률	30	24
화학	3	3	공업	19	25
연영	5	4	약학	17	26
음악	2	5	가톨릭	30	27
식물	4	6	생물	21	28
컴퓨터	7	7	지리	21	29
교통	27	8	천문	33	29
의학	12	9	통신	37	31
군사	11	10	수공	21	32
전기	21	11	예술	26	33
미술	13	12	역사	30	33
동물	6	13	출판	37	33
물리	9	14	교육	43	33
기계	19	15	언론	33	37
문학	8	16	지명	33	37
사회	17	17	항공	37	37
경제	15	18	고적	33	40
광업	21	18	종교	37	40
기독교	26	20	정치	37	42
철학	15	21	책명	37	42
심리	26	22	불교	43	44

4. 결 론

인간사회에서 수수되는 정보의 외양은 이미지·소리·텍스트 등 다양한 모습을 가지고 있지만, 정보축적 및 전달의 가장 보편적인 수단은 자연언어(natural language)이다. 따라서 언어에 대한 이해는 정보의 효율적 처리를 위해서 필수적인 것이라 할 수 있다. 결국 이제 언어는 언어학자만의 연구 대상이 아니라 정보의 효율적 축적과 활용을 다루는 정보과학(information science)의 중요한 연구대상이 된 것이다. 서상규와 한영균(1999)에 기초하여 부언하면, 언어 연구가 더 이상 언어학자의 고유 영역이 아닌 시대, 사람들이 읽고 쓰고 말하는 것이 그대로 연구의 대상이 되는 시대, 무엇보다도 수천만 혹은 수억 어절에 달하는 대량의 자료를 컴퓨터를 이용해서 처리하고 그를 바탕으로 언어(지식)의 내적 구조를 밝혀야 하는 시대, 이것이 오늘날의 언어 연구가 당면하고 있는 시대적 특성이다. 이러한 시대적 특성이 낳은 필연적인 결과가 바로 '계량언어학'이다.

언어의 계량적 성질은 상당히 오래 전부터 주목되어 왔는데, 1830년대에 고안된 모르스 부호에는 이미 문자의 출현 빈도라는 생각이 쓰이고 있다. 그러나 근대적인 통

계학, 통계적 결정 이론이 언어의 성질 연구에 도입된 것은 비교적 최근의 일로서, 1950년대에 들어서면서 부터이다.

한편 앞에서도 언급하였듯이, 이와 같은 계량언어학 연구에서 무엇보다도 중요한 세 가지 축은, '말뭉치'와 이를 가공처리하기 위한 '컴퓨터' 그리고 추출된 언어정보의 '통계분석'이다. 따라서 계량언어학 연구의 활성화를 위해서는, 이 과정에 통계학자의 보다 주도적이고 적극적인 역할이 필요하다고 여겨진다. 그런데 아직까지 계량언어학을 위한 통계적인 방법의 개발이나 연구가 미미한 실정이다. 이에 본 연구에서는 계량언어학 연구 분야에 통계관련 학자들이 더욱 적극적인 관심을 보일 필요성을 강조함과 함께, 예를 통하여 통계적 방법을 활용함으로써 언어연구의 해석을 보다 객관적이고 과학적으로 할 수 있음을 보였다. 본 연구를 출발점으로 하여, 학제간 연구인 계량언어학 연구에 통계학자들이 보다 적극적인 관심을 갖는 계기가 되기를 기대한다.

참고 문헌

1. 강범모 (1999). 빈도와 언어기술, *언어정보의 탐구1*, 연세대학교 언어정보개발연구원.
2. 宮澤達夫 (1970). 어휘의 유사도, *국어학*, 70.
3. 김한샘 (2005). 소설 말뭉치에 나타난 외래어 사용 양상 분석, Corpus Linguistics for Korean Language Learning and Teaching, NFLRC University of Hawaii, 193-236.
4. 남윤진 (2000). *현대국어의 조사에 대한 계량언어학적 연구*, 태학사.
5. 미즈노 슌페이 (2003). 만엽집 권 14 향가 표기에 대한 계량적 고찰, *계량언어학 2집*, 박이정.
6. 박병선 (2005). *한국어 계량적 연구방법론*, 역락
7. 서상규, 한영균 (1999). *국어정보학입문*, 태학사
8. 양경숙 (2003). 코퍼스 규모에 따른 타입과 토큰의 상관성 연구, *제15회 한글 및 한국어 정보처리 학술대회 발표집*, 81-85, 한국.
9. 이상억 (2003). *계량언어학 2집*, 박이정
10. 이정란 (2005). 한국어 학습자언어에 나타난 '-어서'와 '-니까'의 변이 연구, *한국어 학습자의 중간언어 연구*, 커뮤니케이션북스.
11. 임칠성 (2003). 5·18항쟁 관련 유인물과 성명서 어휘의 계량연구(1), *계량언어학 2집*, 박이정.
12. 전혜영 (2005). 연어 구성에 나타난 남녀 은유의 양상, 여성학논집, 22(1), 53-77.
13. 황용주, 최경호 (2006). 신소설의 문법적 연어구성의 국어정보학적 연구, *2006년 어문연구학회 학술대회 발표집*, 한국.
14. 허명회 (1998). 통계적 방법에 의한 한국어 텍스트 유형 및 문체 분석, *언어학*, 22, 3-57.
15. 홍종선 (2003). 계량화에 기반한 국어의 연어 관계 연구, *계량언어학 2집*, 박이정.
16. 황용주 (2006). 신소설의 연어연구, 제47회 *한국언어문학회 학술대회*

발표집, 한국.
17. 황용주 (2007). *연어 구성의 계량언어학적 연구*, 전북대학교 박사학위논문.
18. Barnbrook, G. (1996). *Language and Computer*, Edinburgh University Press.
19. Kilgariff, A. (1996). Which words are particularly characteristic of text? A survey of statistical approach, *Short Reports*, Information Technology Research Institute, University of Brighton, 6 March.
20. McEnery, T., Wilson, A. (1998). *Statistics for corpus linguistics*, Edinburgh University Press.

[2007년 1월 접수, 2007년 3월 채택]