Journal of the Korean Data & Information Science Society
2011, **22**(1), 49–55

한국데이터정보과학회지

통계적 방법을 활용한 객관적 언어정보 도출 - 학제적 연구의 가능성 모색

최경호[1] · 이용욱[2]

[1]전주대학교 기초의과학과 · [2]전주대학교 한국어문학전공

접수 2010년 11월 06일, 수정 2010년 12월 21일, 게재확정 2011년 01월 12일

요 약

최근 들어 여러 영역에서 통섭을 통한 융합을 이루려는 시도가 자주 나타난다. 학문에 있어서도 예외는 아닌바, 학제적 연구가 그 예라 하겠다. 통계학과 관련된 학제적 연구의 한 분야로 언어정보학 또는 계량언어학이라 일컬어지는 언어학 연구가 있다. 그런데 통계학과 언어학의 학제적 연구는 주로 언어학자들을 중심으로 이루어져 오고 있다. 따라서 통계학적인 측면에서 보면 언어학자들의 연구결과에 일부 부족한 부분이 분명 존재한다. 이에 본 연구에서는 일부 언어학 연구에서 나타나는 객관성 확보의 부족한 면에 대한 보완을 통계적인 방법을 이용하여 수행함으로써, 통계학과 언어학의 '학제적 연구'의 완성도를 높일 수 있는 방안에 대해 고찰해 보았다. 부언하면 본 연구에서는 언어학 연구에서 보다 객관적인 언어정보를 도출하는데 도움이 될 수 있는 여러 통계적인 방법들을 소개하고 응용 예를 보였다.

주요용어: 계량언어학, 사회네트워크분석, 언어정보, 통계적 방법.

1. 서 론

21세기 들어 많은 학자들에 의해서 자주 거론되는 용어중의 하나가 '통섭 (consilience)'이다. 우리사회에서 '통섭'이라는 용어가 널리 주목받게 된 까닭은 아마도 우리사회의 학문 간 벽이 매우 높고 폐쇄적인 데 대한 비판적 관심 때문인 것으로 보인다.

불과 최근까지만 하더라고 우리나라에서는 영역의 세분화가 전문화라는 미명하에 일반적 현상으로 자리 잡고 있었던 것이다. 세분화(전문화)는 또한 변화와 창의적 발상을 저해하는 요인으로 작용하기도 한다. 새로운 이념이나 기술개발은 다른 영역의 지식을 빌리지 않고서는 이루어지기 어려운 것이다. 인간이 사회 속에서 다른 사람들과의 상호 작용을 통하여 성장하고 발전하듯이 어느 한 분야의 변화와 발전도 다른 영역과의 상호 작용 없이 이루어질 수는 없는 것이다 (박행렬, 2009).

이러한 현상은 학문에 있어서도 예외는 아니었는데, 학문에 있어서 이러한 문제점을 해결하기 위해 제시된 방법이 바로 '학제적 연구 (interdisciplinary research)' 방식이다. 학제적 연구 방식은 대학사회에서는 학제적 연구소 설립, 학과 통폐합과 학부제 도입, 연계전공 및 통합과정 도입, 팀티칭 등으로 나타났다 (안상헌, 2009). 물론 '통섭'과 '학제적 연구'가 동일한 개념인지에 대해서는 이견이 있지만, '융합'을 통하여 세분화가 가진 문제점을 해결하고자 하는 지향점에서는 큰 차이가 없다고 하겠다.

[1] 교신저자: (560-759) 전북 완산구 효자동 3가 1200, 전주대학교 기초의과학과(통계학), 교수.
E-mail : ckh414@jj.ac.kr
[2] (560-759) 전북 완산구 효자동 3가 1200, 전주대학교 한국어문학 전공, 조교수.

최근 들어 언어학자들에 의하여 통계학 영역과의 학제적 연구를 시도한 사례들이 나타났는데 이는 시사하는 바가 크다. 구체적으로 살펴보면 박병선 (2005)의 '한국어 계량적 연구방법론'과 황용주 (2007)의 '연어 구성의 계량언어학적 연구' 그리고 신효필 (2009)의 '언어학과 통계모델' 및 김동성 (2010)의 '언어자료분석을 위한 통계학' 등이다. 이들 연구는 모두 언어학자들이 그들의 관점에서 필요한 통계영역을 다루고 있지만, 통계학이 한 축이 되는 '학제적 연구'의 좋은 보기라 할 수 있다.

이러한 시점에서 일부 연구에서 나타나는 객관성 확보의 부족한 면에 대한 보완을 함으로써, 통계학과 언어학의 '학제적 연구'의 완성도를 높일 수 있는 방안에 대한 고찰은 의미가 있다고 사료된다. 이에 본 연구에서는 언어학 연구에서 보다 객관적인 언어정보를 도출하는데 도움이 될 수 있는 여러 통계적인 방법들을 소개하고자 한다. 본 연구를 위하여 귀한 자료를 제공해주신 고려대학교 언어정보연구소 김동성 교수께 감사를 드린다.

2. 통계학과 언어학

언어학이란 인간의 언어와 관련한 여러 현상을 과학적인 방법으로 연구하는 학문으로 언어의 기능과 본질, 언어의 역사, 언어의 변이, 언어와 인간관계 따위를 주로 연구하는 분야이다. 이러한 언어학 연구에서 그동안 사용되어 온 일반적인 연구 방법은, 모국어 화자의 직관에 의존한 논리적 설명이 대부분이었다. 그러나 최근 연구는 직관에 의존한 연구에서 대규모의 실제적 자료를 이용하는 연구로 옮아가고 있다. 그런데 이 과정에서 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). 이를 위하여 탄생된 연구 분야가 바로 대규모 자료처리에 필수적인 통계적인 처리방법론 등을 언어연구에 복합적으로 활용한 계량언어학 (quantitative linguistics)이다.

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). 이에 통계학이 주로 이용되는 언어학 연구 분야가 바로 계량언어학인 바, 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 2.1과 같이 나타내고 있다 (최경호와 황용주, 2007).

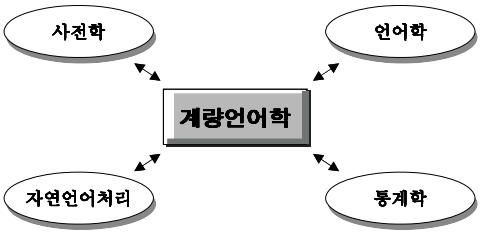

그림 2.1 계량언어학과 인접 학문과의 연계관계

3. 언어학연구에서 통계적 방법의 활용

본 연구는 언어학과 통계학의 학제적 연구의 활성화에 기여하고자, 기존 언어정보 도출을 위한 언어학 연구에서 보다 객관화된 결론을 도출하는데 통계학이 기여할 수 방법을 소개하고자 하는데 그 목적이 있다. 이 목적을 달성하기 위해 먼저 기 수행된 언어학 연구 몇 가지를 먼저 소개해 보면 다음과 같다.

정한진 등 (2007)은 성인화자의 어휘 (동사) 이름대기 능력의 비교 준거 마련과 임상적인 평가 및 치료의 활용에 도움을 주고자, 코퍼스를 이용한 사용빈도가 높은 어휘와 기초 어휘에 대한 연구들과 친밀도를 토대로 성인용 동사 이름대기 어휘목록을 구성하고 제시하였다. 이들이 실험연구에서 활용한 84개 어휘목록은 표 3.1과 같다.

표 3.1 정한진 등 (2007)이 사용한 성인용 동사이름대기 어휘목록

가다	가르치다	걸다	결혼하다	고르다	고부하다	그리다	기다리다
끌다	낚시하다	낳다	내리다	넘다	넣다	놀다	놓치다
누르다	다치다	닦다	달리다	던지다	도착하다	돌리다	돕다
듣다	떨어지다	뜨다	만나다	만들다	맞다	먹다	밀다
받다	버리다	벗다	보다	보이다	부르다	불다	빼다
뿌리다	사다	사용하다	생각하다	서다	소개하다	싣다	심다
싸우다	쌓이다	쓰다	씻다	앉다	열다	열리다	오다
올리다	요리하다	웃다	일하다	읽다	입다	잇다	잊다
자다	자라다	자르다	잡다	잡히다	저축하다	전화하다	주다
지적하다	지키다	찍다	찾다	치다	치료하다	타다	팔다
풀다	피우다	흐르다	흘리다				

표 3.1의 어휘목록을 이용하여 51명의 성인을 대상으로 일상생활에서 실제로 사용하는 빈도와 사용정도와 관계없이 평소에 친밀하게 느끼는 정도를 5점 척도로 측정하고 점수를 계산하여 분류표로 정리한 후 친밀도와 실제사용빈도 간에 많은 차이를 보이고 있음을 주장하였다. 그런데 이 과정에서 이들은 통계적 방법을 통한 객관적인 정보에 기초하지 않고 주관적인 판단에 기초하여 결론을 내림으로써 객관성을 상실하고 있다. 나아가 주장하는 '차이'가 어떤 의미인지 명확하지 않다. 측정된 점수 간의 차이인지, 84개 어휘의 사용빈도 순위 간의 차이를 의미하는지가 명확하지 않다. 만약 측정된 점수 간의 차이라면 대응검정 (paired t-test)과 같은 통계적인 방법을 활용하여 친밀도 점수와 실제사용빈도 점수 간에 차이가 있음을 보인다면 주장이 보다 객관적으로 입증될 수 있을 것이다. 실제로 정한진 등 (2007)에 주어진 자료를 2차 자료로 활용하여 84개 어휘에 대한 친밀도와 실제사용빈도 점수 간의 대응분석을 활용한 평균유의차 검정을 수행해 보면 유의확률<0.001로 통계적으로 유의한 차이를 보인다. 따라서 코퍼스로부터 추출된 84개의 고빈도 어휘의 친밀도 점수와 실제사용빈도 점수 간에는 차이가 있음을 주장할 수 있다. 한편 만약 주장하고자하는 차이가 사용빈도의 순위 간 차이라면 순위상관계수를 구하고, 이를

활용하여 결론을 유도해야 한다. 예컨대 이 경우 친밀도와 실제사용빈도에 대하여 순위상관계수를 구해보면 0.958로 유의확률이 0.0001보다 작아 통계적으로 유의하게 나타난다. 따라서 '고빈도라고 평가된 몇몇 어휘를 제외하고 주관적인 빈도와 많이 달랐다'라고 언급하면서 '연구의 결과가 외국의 경우와는 다른 결과를 보였다'는 주장에는 문제가 있게 된다. 어느 측면을 보이고자 하는지는 몰라도 통계적인 방법을 사용한다면 더욱 객관적인 결론에 도달할 수 있을 것으로 판단된다.

한편 김동성 (2009)은 '언어치료를 위한 어휘들의 공간모델'에서 정한진 등 (2007)의 연구에서 친숙도와 개인적 사용빈도 성향에서 고르게 분포한 어휘 26개를 골라 세종 1,500만 어절 코퍼스를 토대로 정한진 등 (2007)이 갖는 언어학적인 문제점을 지적하였다. 즉 김동성 (2009)은 정한진 등 (2007)의 연구에서 제시된 빈도가 코퍼스로부터 도출된 출현빈도와 매우 다름을 보이고, 이는 어휘들 간에 간섭현상이 있는 바 응답과정에서 간섭현상이 발생된 결과일 수도 있다고 하였다. 나아가 언어치료용 어휘목록 작성을 위한 해결방안의 하나로 어휘들 간의 벡터 의미 공간 모델을 제시하고 공간적 연관성을 조사하였다. 이를 위하여 26개 어휘 각각을 중심어로 하여 좌우 50어절씩을 대상으로 형태소분석을 통하여 상대어휘가 포함된 빈도수를 계산하고 이를 표로 제시하였다. 나아가 이 표를 이용하여 벡터 연산을 통한 상대적 출현빈도를 계산하고 그 결과가 정한진 등 (2007)과는 다르지만 1,500만 전체 코퍼스에서 발견되는 순서와는 유사함을 보임으로써 정한진 등 (2007)의 연구가 갖는 문제점을 지적하였다. 그런데 김동성 (2009)의 연구에 있어서도 몇 가지 아쉬운 점이 있다. 예컨대 어렵게 정리한 표의 정보를 상대적 출현빈도를 계산하는데 활용하는데 그칠 것이 아니라, 그림 3.1과 같이 유사성 측도에 기초한 군집을 이용하여 시각화해서 보여준다면 더욱 독자의 이해를 높일 수 있을 것이다. 나아가 빈도만을 이용할 경우에 비하여 보다 의미 있는 정보 도출도 가능할 것으로 여겨진다. 그림 3.1로부터 왕래발착 (往來發着) 동사인 '오다'와 '가다'가 먼저 묶이고 다음으로 '사다'와 '먹다'가 그룹화 됨을 알 수 있다. 참고로 그림 3.2는 사회연결망을 통한 일부 어휘들 간의 공기 (co-occurrence)정도를 알 수 있는 그림이다.

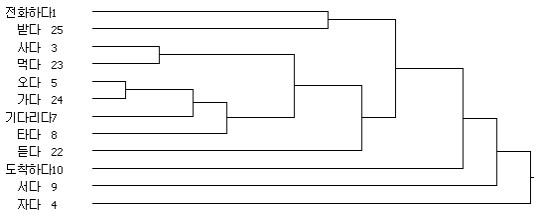

그림 3.1 덴드로그램(일부)

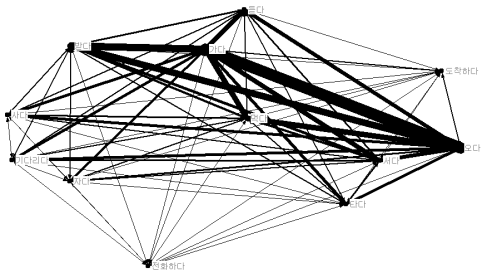

그림 3.2 연결망분석을 통한 어휘 간 관계도

다음으로 통계적 검정을 활용함으로써 도출된 언어정보에 대하여 보다 객관적인 결론을 유도할 수 있다. 예를 들어, 황용주 (2007)는 신소설과 현대소설에 대한 코퍼스를 구축하고 '담배'에 대한 연어(collocation)를 비교·연구하였다. 이를 위하여 식 (2.1)과 (2.2) 등의 z-값이나 t-값을 이용하였다. 그런데 강범모 (2003)에 따르면 이들을 활용한 계산은 간단하면서도 신속하게 계산할 수 있는 장점이 있지만, 전체적으로 단어의 유의성을 크게 만드는 경향이 있다.

$$z = \frac{O - E}{\sigma} \tag{3.1}$$

$$t = \frac{O - E}{\sqrt{O}} \tag{3.2}$$

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

따라서 이런 경우에는 Manning과 Schutze (1999)가 제안한 t-분포를 이용한 검정이 더욱 바람직하다. 이론적으로 보면 w_1과 w_2가 연어라는 것은 두 단어가 출현하는 것이 우연이 아니라는 것과 같다. 반대로 귀무가설은 w_1과 w_2가 연어가 아니므로 우연한 출현이다 (박희창, 2010). 따라서 두 단어는 독립적으로 출현하게 되며, 두 단어의 출현 확률도 개별 단어의 독립확률의 연산과 같다 (김동성, 2010). 이에 106만 어절을 대상으로 한 신소설에서 '빨다'의 빈도와 '담배'의 빈도는 각각 20번과 138번이다. 따라서 P(빨다 ∩ 담배)은 다음과 같다.

$$P(\text{빨다} \cap \text{담배}) = P(\text{빨다}) \times P(\text{담배}) = \frac{20}{1,060,000} \times \frac{138}{1,060,000} = 0.000000002$$

한편 '담배'와 '빨다'가 동시에 출현한 빈도수는 6회이므로

$$P(\text{담배 빨다}) = \frac{6}{1,060,000} = 0.00000566$$

이다. 따라서

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \left(= \frac{\widehat{p} - p_o}{\sqrt{\widehat{p}(1-\widehat{p})/n}} \right) = \frac{0.00000566 - 0.000000002}{\sqrt{0.00000566/1060000}} = 2.44$$

로 유의수준 1%에서 귀무가설이 채택되어 '담배'와 '빨다'는 연어라 할 수 없다. 이에 반하여 1,003,071어절을 대상으로 한 현대소설의 경우에는

$$P(\text{빨다} \cap \text{담배}) = \frac{75}{1,003,071} \times \frac{680}{1,003,071} = 0.00000005,$$

$$P(\text{담배 빨다}) = \frac{44}{1,003,071} = 0.0000438$$

이다. 따라서 $\mathrm{t} = (\bar{x} - \mu)/\sqrt{s^2/n} = 6.621$로 유의수준 1%에서 귀무가설이 기각되어 '담배'와 '빨다'는 연어라 할 수 있다. 이렇듯 통계적 검정을 활용하면 연어 관계를 보이는데 있어서도 보다 객관적인 입증이 가능하게 된다.

4. 결 론

지난 20세기는 전문화의 이념이 지배한 시대였다고 볼 수 있다. 모든 분야가 세분화되고 그 세분화된 영역 속에서 깊이 파고드는 노력만이 인정을 받았으며, 제너럴리스트는 별 주목을 받지 못했던 시대였던 것이다. 그러나 이러한 세분화·전문화는 영역 상호간에 높은 벽을 쌓아 자기 이외의 분야에 대해 무지한 결과를 낳았고, 사물에 대한 단일한 인식체계를 갖지 못하게 하는 원인으로 작용하였다. 이러한 한계적 상황을 극복하고자 나온 개념이 바로 통섭 및 융합의 개념이다 (박행렬, 2009).

통섭은 학문에 있어서도 예외 없이 나타났는데, 학제적 연구가 바로 그것이다. 언어학을 예로 들면 그동안에는 국어학자들을 중심으로 연구가 수행되었지만, 서상규와 한영균 (1999)이 언급한 바와 같이 언어 연구가 더 이상 언어학자의 고유 영역이 아닌 시대 즉 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 결과 중의 하나가 바로 '계량언어학'이다.

계량언어학 연구에서 무엇보다도 중요한 세 가지 축은, '코퍼스'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 최경호 (2007)나 차경엽과 이성덕 (2008) 등의 연구가 있기는 하지만, 아직까지 계량언어학을 위한 통계적인 방법의 개발이나 연구가 미미한 실정이다. 이에 본 연구에서는 계량언어학 연구 분야에 통계관련 학자들이 더욱 적극적인 관심을 보일 필요성 알리고자, 통계적 방법을 활용함으로써 언어정보의 도출 및 해석을 보다 객관적이고 과학적으로 할 수 있음을 예를 통하여 보였다. 향후 본 연구를 기와로 통계학을 중심으로 한 학제적 연구가 더욱 활성화되고, 이점이 어려운 여건에 처한 지방대학 통계학 관련 교수들의 활로를 모색하는 한 방편이 되기를 희망해 본다.

참고문헌

강범모 (2003). <언어, 컴퓨터, 코퍼스 언어학>, 고려대학교 출판부, 서울.

김동성 (2009). <인문치료의 모색>, 강원대학교 인문과학연구소, 강원.

김동성 (2010). <언어자료분석을 위한 통계학>, 한국외국어대학교 출판부, 서울.

박병선 (2005). <한국어 계량적 연구방법론>, 역락, 서울.

박행렬 (2009). 한국 법과학의 영역 간 통섭에 관한 연구. <한국공안행정학회보>, **34**, 123-156.

박희창 (2010). 올바른 연관성 규칙 생성을 위한 의사결정과정의 제안. <한국데이터정보과 학회지>, **21**, 263-270.

서상규, 한영균 (1999). <국어정보학입문>, 태학사, 서울.

신효필 (2009). <언어학과 통계모델>, 서울대학교출판문화원, 서울.

안상현 (2009). 사회생물학적 '통섭'의 이데올로기적 성격. <인문학지>, **38**, 155-182.

임칠성 (2003). <계량언어학2>, 박이정, 서울.

정한진, 이옥분, 서경희 (2007). 성인용 동사 이름대기 평가 어휘 목록. <언어치료연구>, **16**, 161-172.

차경엽, 이성덕 (2008). 실용통계 개발을 위한 새로운 제안. <한국데이터정보과학회지>, **19**, 187-195.

최경호 (2007). 문학작품에 대한 계량언어학적 분석. <한국데이터정보과학회지>, **18**, 1057-1064.

최경호, 황용주 (2007). 계량언어학 연구에서 통계적 방법의 활용. <한국데이터정보과학회지>, **18**, 269-278.

황용주 (2007). <연어 구성의 계량언어학적 연구>, 박사학위논문, 전북대학교, 전북.

Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*, MIT Press, Cambridge, Massachusetts.

Journal of the Korean Data & Information Science Society
2011, **22**(1), 49–55

한국데이터정보과학회지

The deduction of objective linguistic information using statistical methods - The grouping of the possibility of interdisciplinary research

Kyoung Ho Choi[1] · Yong Wook Lee[2]

[1]Department of Basic Medical Science, Jeonju University
[2]Korean Language and Literature Major, Jeonju University

Received 06 November 2010, revised 21 December 2010, accepted 12 January 2011

Abstract

There are tries to unite through consilience in many fields. Interdisciplinary research is an instance of those. Linguistic studies called linguistic informatics or quantitative linguistics is a field of interdisciplinary research related with statistics linguists have studied chiefly statistics and linguistics. In the statistical aspect, there is need to supplement somewhat of the result of researches by linguists. This study shows statistical method can supplement insufficient objectivity in linguistic studies, and examines the way to raise a degree of completion of interdisciplinary research on statistics and linguistics. This study also shows an introduction and application of the statistical method can be useful for the deduction of objective linguistic information in linguistic studies.

Keywords: Linguistic information, quantitative linguistics, social network analysis, statistical methods.

[1] Corresponging author: Professor, Dept. of Basic Medical Science, Jeonju University, 560-759, Korea. E-mail : ckh414@jj.ac.kr
[2] Assistant Professor, Korean Language and Literature Major, Jeonju University, 560-759, Korea.