언어 네트워크 분석 방법을 활용한 학술논문의 내용분석*

A Content Analysis of Journal Articles Using the Language Network Analysis Methods

이수상 (Soo-Sang Lee)**

초 록

본 연구의 목적은 국내 학술논문 데이터베이스에서 검색한 언어 네트워크 분석 관련 53편의 국내 학술논문들을 대상으로 하는 내용분석을 통해, 언어 네트워크 분석 방법의 기초적인 체계를 파악하기 위한 것이다. 내용분석의 범주는 분석대상의 언어 텍스트 유형, 키워드 선정 방법, 동시출현관계의 파악 방법, 네트워크의 구성 방법, 네트워크 분석도구와 분석지표의 유형이다. 분석결과로 나타난 주요 특성은 다음과 같다. 첫째, 학술논문과 인터뷰 자료를 분석대상의 언어 텍스트로 많이 사용하고 있다. 둘째, 키워드는 주로 텍스트의 본문에서 추출한 단어의 출현빈도를 사용하여 선정하고 있다. 셋째, 키워드 간 관계의 파악은 거의 동시출현빈도를 사용하고 있다. 넷째, 언어 네트워크는 단수의 네트워크보다 복수의 네트워크를 구성하고 있다. 다섯째, 네트워크 분석을 위해 NetMiner, UCINET/NetDraw, NodeXL, Pajek 등을 사용하고 있다. 여섯째, 밀도, 중심성, 하위 네트워크 등 다양한 분석지표들을 사용하고 있다. 이러한 특성들은 언어 네트워크 분석 방법의 기초적인 체계를 구성하는 데 활용할 수 있을 것이다.

ABSTRACT

The purpose of this study is to perform content analysis of research articles using the language network analysis method in Korea and catch the basic point of the language network analysis method. Six analytical categories are used for content analysis: types of language text, methods of keyword selection, methods of forming co-occurrence relation, methods of constructing network, network analytic tools and indexes. From the results of content analysis, this study found out various features as follows. The major types of language text are research articles and interview texts. The keywords were selected from words which are extracted from text content. To form co-occurrence relation between keywords, there use the co-occurrence count. The constructed networks are multiple-type networks rather than single-type ones. The network analytic tools such as NetMiner, UCINET/NetDraw, NodeXL, Pajek are used. The major analytic indexes are including density, centralities, sub-networks, etc. These features can be used to form the basis of the language network analysis method.

키워드: 언어 텍스트, 언어 텍스트 분석, 언어 네트워크, 언어 네트워크 분석, 내용분석
language text, language text analysis, language network, language network analysis, content analysis

---

* 이 논문은 2013년도 부산대학교 인문사회연구기금의 지원을 받아 연구되었음.

** 부산대학교 문헌정보학과 교수(sslee@pusan.ac.kr)

■ 논문접수일자 : 2014년 11월 19일 ■ 최초심사일자 : 2014년 11월 26일 ■ 게재확정일자 : 2014년 12월 12일
■ 정보관리학회지, 31(4), 49-68, 2014. 〔http://dx.doi.org/10.3743/KOSIM.2014.31.4.049〕

1. 서 론

언어로 된 텍스트를 네트워크 분석 대상으로 하여, 그 내용을 분석하는 방법을 언어 네트워크 분석(language network analysis)이라고 한다. 언어 텍스트로 표현된 메시지에 내재된 다양한 특성들을 나타내는 개념들을 추출하고, 그들 간에 형성되는 의미적 관계의 속성들을 파악하고자 할 때 언어 네트워크 분석 방법을 사용하면 매우 유용하다. 일반적으로 언어 텍스트의 특성을 나타내는 개념은 키워드(또는 단어)로 표현되며, 명사형태의 단어, 특정한 범주에 속하는 단어, 감성을 나타내는 단어 등으로 나타난다.

방법론으로 보면, 언어 네트워크 분석은 내용분석(content analysis) 방법의 범주에 해당된다고 볼 수 있다. 전통적인 내용분석은 연구논문, 언론기사, 인터뷰자료, 기록자료 등과 같은 언어 텍스트에서 특정한 개념들(저자, 년도, 주제 등의 특성)이 등장하는 경향을 빈도와 같은 통계적 데이터로 파악하는 방법이다. 반면에, 언어 네트워크 분석은 언어 텍스트로부터 특정한 개념들의 관계를 파악하고, 이것을 네트워크로 구성하여, 계량적인 특성을 분석하는 것까지 확대된 내용분석 방법이다. 이러한 개념들 간의 관계를 언어 네트워크(language network)로 표현한다.

그동안 언어 네트워크 분석 방법은 문헌정보학에서 계량정보분석 방법의 한 영역인 동시단어분석, 언어학이나 심리학에서 언어의 의미적 특성을 분석하는 방법, 인터뷰 내용의 질적자료를 분석하는 연구방법 등에서 사용하여 왔다. 서로 다른 목적과 관점에서 사용하고 있지만, 언어 네트워크 분석을 수행하는 구체적인 절차는 매우 유사한 형태를 나타내고 있다. 물론 사용하는 용어에 있어서는 서로 간에 차이를 나타내기도 한다. 즉 언어 네트워크를 의미(semantic) 네트워크, 개념(concept) 네트워크, 단어(word) 네트워크, 키워드(keyword) 네트워크, 네트워크 텍스트(network text) 등으로도 표현하고 있지만(박치성, 정지원, 2013), 이 네트워크들을 분석하는 과정에서 유사성이 아주 많다는 것이다.

국내에서는 언어 네트워크라는 용어를 가장 많이 사용하고 있지만, 내용적으로는 의미 네트워크가 더 잘 어울린다고 할 수 있다. 대체로 영어로는 'semantic network'라 표기하며, 한글로는 '언어 네트워크'라 부르고 있다. 굳이 구분한다면, 언어 네트워크 분석은 언어 텍스트 자체의 관계적 분석이라는 점을 강조하는 표현이고, 의미 네트워크 분석은 언어 텍스트에 내재된 의미론적 속성에 대한 언어학적 분석이라는 점을 강조하는 표현이라 할 수 있다.

최근 들어 문헌정보학을 포함하여 다양한 주제 영역에서 언어 네트워크 분석 방법을 적용한 학술논문들이 많이 등장하고 있다. 절차적 유사성에도 불구하고, 주제 영역들마다 조금씩 차이가 나는 개념들과 기법들을 사용하고 있다. 따라서 언어 네트워크 분석 관련 학술논문들을 수집하고, 하나의 방법론적인 체계로서 그 방법론적인 특징의 내용을 세밀하게 분석하여, 그것의 유사점과 차이점을 확인하고, 그것들을 포괄하는 언어 네트워크 분석의 새로운 절차를 정리할 필요가 있다. 이 작업은 특정한 학문영역에서 사용하는 개별적인 분석방법이 아니라, 언어 텍스트의 내용을 분석하고자 하

는 모든 주제영역을 포괄하는, 보다 다학문적 관점의 연구방법으로서 언어 네트워크 분석 방법론의 역할을 정립하는데 기초가 될 수 있기 때문이다.

이러한 연구목적에 따라 본 연구는 KCI, RISS, DBPIA, KISS 등 국내 학술논문 데이터베이스에서 언어 네트워크 분석과 관련된 학술논문들을 검색하여, 방법론으로서 요구되는 몇 가지 범주에 따라 미시적인 내용분석을 시도하게 된다. 내용분석 대상은 국내학술논문으로 한정하여 검색된 전체 53편이다. 그리고 내용분석의 범주는 분석대상의 언어 텍스트 유형, 키워드 선정 방법, 동시출현관계의 파악 방법, 네트워크의 구성 방법, 네트워크 분석도구, 그리고 분석지표의 유형과 같이 6가지이다.

2. 언어 네트워크 분석

2.1 언어 텍스트의 분석 방법

우리 주변에는 다양한 유형의 언어 텍스트가 존재한다. 그동안 각종 연구논문이나 언론기사 등에서 분석의 주요한 대상이 되어왔던 언어 텍스트는 크게 심층면담, 회의나 토론 현장에서 녹취한 정성적 텍스트 자료(면담자료, 토론자론 등)이거나 이미 형식적인 체계를 갖추어 발표된 문헌적 텍스트 자료(언론기사, 연구문헌, 기록자료 등)로 구분된다. 어떠한 유형의 텍스트이든지 이를 분석하기 위해 사용할 수 있는 주요 방법은 크게 정성분석 방법, 내용분석 방법, 언어분석 방법, 언어 네트워크 분석 방법으로 구분할 수 있다.

첫째, 전통적인 정성분석은 대부분 응답자의 인터뷰 또는 집단토론에서 얻어진 것을 분석자가 스크립트화된 텍스트로 변환하고, 그것의 내용 중에서 특정한 범주에 해당되는 내용들만을 취합하여 해석하는 방식으로 진행한다. 정성자료의 단순한 기술이 아니라 해석에 집중하기 때문에, 이 방법은 분석자의 주관이나 가치관에 좌우될 수 있고, 동일한 텍스트를 사용하더라도 분석자마다 다른 결과를 나타낼 우려가 있다. 연구방법의 관점에서 보면, 전통적인 질적연구 방법에 해당된다.

둘째, 자동화된 정성분석은 전통적인 정성분석을 S/W 도구를 사용하여 수행하는 방법이다. 대표적인 S/W 도구는 NVivo이다. NVivo는 정성분석의 이론적 기반이 되는 근거이론을 바탕으로 개발된 프로그램으로써, 화자의 관점에 의거하여 자료의 범주화 및 조직화를 컴퓨터 S/W의 도움으로 처리할 수 있다. 근거이론을 바탕으로 한다는 의미는 연구자의 관점이 아닌 원자료에 의해 실제 현상을 설명할 수 있도록, 원자료로부터 상위범주로 가는 개방코딩(open-coding)을 사용한다는 의미이다. 전통적인 정성분석 방법과 비교해보면, 수집된 자료를 대상으로 S/W를 이용하여 범주를 코딩하고 단계적으로 조직하여 분석함으로써 연구자의 주관적 판단을 줄여서 연구의 타당성과 신뢰성을 높일 수 있다는 장점이 있다.

셋째, 내용분석은 정성분석의 오류에 대한 대안적인 분석방법이며, 메시지의 내용을 일정한 분류범주에 따라 코딩한 다음, 그 결과를 계량적으로 취합한다. 정성 데이터의 계량분석에 해당되며, 계량화된 데이터는 빈도분석, 상관분석 등과 같은 적절한 수준의 통계분석을 거쳐

해석할 수도 있다. 코더(coder)의 수준에 따라 코딩의 결과가 다를 수 있고, 대용량의 데이터인 경우 시간과 비용이 소요될 수 있는 단점이 있다. 아무튼 전통적인 내용분석 방법은 분석대상의 언어 텍스트 집합에서 각 메시지의 주요 특성을 파악하여 전체의 관점에서 정리하는 메타분석이라 할 수 있다. 그래서 전체 메시지에 나타난 추세, 경향, 핵심 논제 등을 파악하는데 유용하다. 그리고 특정 메시지의 의미분석, 특정 메시지의 시간적 추이분석, 매체 간 메시지의 차이분석 등도 가능하다.

넷째, 언어분석은 메시지에서 특정한 의미를 가지는 단어들을 추출하여 그들이 출현하는 특성을 분석하는 방법으로, LIWC(Linguistic Inquiry and Word Count)와 같은 분석도구(한국의 경우 K-LIWC 사용)를 활용한다(이창환, 심정미, 윤애선, 2005). 현재까지 활용하고 있는 언어분석은 사전에 프로그램에서 정의된 기준에 해당하는 단어들(주로 언어학적/심리적 변인을 나타내는 단어)의 출현 현황만을 알 수 있으며, 단어들에 대한 그 이외의 분석이 가능하지 않다. 그리고 언어 텍스트에서 형태소 단위로 추출한 개별 단어들에 대한 출현 비율만을 제시하므로, 이 단어들이 모여 궁극적으로 어떠한 의미를 나타내고 있으며, 단어들 간에 어떠한 관계가 형성되는지를 알 수 없다는 단점이 있다.

다섯째, 언어 네트워크 분석은 기존 언어 텍스트의 분석 방법들을 통합하면서, 개념들간의 의미적 관계에 나타나는 새로운 특성을 파악하는 방법이 추가된, 보다 복합적인 방법으로서 자리매김을 할 수 있다. 그래서 이 방법은 미시적 관점뿐만 아니라 거시적 관점에서 텍스트에 내재된 다양한 의미를 계량적으로 파악하고 분석하는 방법이 된다.

2.2 언어 네트워크 분석의 이해

언어 네트워크 분석은 언어로 된 메시지(텍스트)에서 의미(개념)를 가지는 단어들을 추출하고, 핵심적인 역할을 하는 단어인 키워드를 부여하며, 언어 메시지 내에서 구성되는 그들의 연결관계를 파악하여 네트워크를 생성하여, 언어 메시지의 다양한 특성을 분석하는 작업을 말한다. 먼저, 언어 네트워크 분석을 정의한 다양한 견해를 정리하면 다음과 같다.

첫째, 텍스트의 의미적 연관관계를 강조하는 견해이다. 텍스트로 표현된 메시지의 내용에 나타난 주요한 개념 사이의 의미론적 연관(semantic association) 관계를 파악하여 텍스트에 내재된 다양한 특징을 파악한다는 입장이다. 구체적으로 설명하면, 문장과 같은 하나의 경계(범위) 내에서 함께 사용된 단어(키워드)의 빈도를 파악하고, 각 단어에 대한 적절한 분석지표가 계산되고, 이를 통해 전체 네트워크에서 특정한 단어가 가지는 의미를 알아보는 것이다. 단어가 사용된 빈도만을 고려하여 의미를 찾고자 했던 전통적인 내용분석과는 달리 특정 단어와 함께 자주 사용되는 단어가 무엇인지를 파악하여 단어 사이의 구조적 관계를 파악할 수 있는 차이가 있다는 견해이다(정덕호, 이준기, 김선은, 박경진, 2013).

둘째, 내용분석 방법이 보다 강조되는 견해이다. 기존의 내용분석 방법의 빈도분석뿐만 아니라 보다 다양한 관계분석 지표들을 사용하여 내용분석의 수준을 높인다는 의미이다. 즉 텍스트 내의 주제 범위에 따라 개념 간의 관계를 확인하기 위해 자주 함께 출현하는 단어의 쌍을 살

펴보는 내용분석 기술이며, 의미있는 단어들의 출현을 파악하여 텍스트 주제 파악을 용이하게 하고, 개념의 유사성이나 근접성을 바탕으로 자주 함께 출현하는 단어들을 군집화하거나 네트워크로 연결하는 방법인 것이다(김혜영, 이도길, 강범모, 2011).

셋째, 네트워크 분석 방법을 강조하는 견해이다. 언어 텍스트의 의미분석에 사회 네트워크 분석방법을 활용한다는 점을 강조한다. 언어 네트워크 분석은 텍스트 내의 단어들 간의 관계를 부호화하고(encoding) 연계된 단어들 간의 네트워크를 구성하는 기법이다(김유호, 2012). 단어 사이의 연결양식을 분석하여 가시화함으로써 추상적인 의미구조를 구체화하는 데 용이하다. 또한 단어들 사이의 관계를 시각적인 네트워크로 묘사하여 중심적 단어와 주변 단어들 사이의 관계가 어떠하며, 어느 정도의 강도로 연결되어 있는지 한 눈에 알아볼 수 있는 장점을 가지고 있다(이혜준, 이동일, 이주현, 2010). 기존의 사회 네트워크 분석 방법을 활용하여 언어 텍스트의 의미를 네트워크로 모델링하여 분석하고자 하는 견해이다.

마지막으로 언어의 인지적 구조 분석과 관련된 견해이다. 언어 네트워크 분석은 인간이 사용한 언어에 대한 인지시스템의 구조와 작동원리를 파악하여 모델화하는 것이라는 주장이다(정석환, 2013). 그리고 인간의 언어에 나타난 단어들 간의 네트워크 구조를 분석함으로써 인간들 간의 인지구조(cognitive structure)의 차이를 확인할 수 있다는 것이다(심준섭, 2012).

종합하면, 언어 네트워크 분석(또는 의미 네트워크 분석)은 언어로 된 텍스트로부터 의미를 나타낼만한 개념을 단어의 형태로 추출하고, 그들 간의 동시출현과 같은 연관관계를 토대로 네트워크를 구성하여, 텍스트의 의미적 내용을 분석하는 네트워크 기반의 내용분석 방법이다. 기존의 내용분석에 의미적 특성의 분석이 추가되고, 네트워크 분석이 결합된 것이라는 것이다. 어떤 언어 텍스트가 나타내는 의미(개념)는 그 텍스트에 등장하는 단어들로 설명이 되고, 한 텍스트에 등장하는 단어들은 그 텍스트의 전체적인 의미를 설명하기 위해 함께 등장하며, 이러한 의미적인 결합 관계를 네트워크로 모델링하면 다양한 특성들이 분석된다. 따라서 언어 네트워크는 언어 텍스트를 의미지도(semantic map)로 재구성한 것이 된다. 여기서 의미 또는 개념을 나타내는 단어는 텍스트에서 추출한 여러 단어들 중에서 선정하는 것이기에 키워드(keyword) 또는 핵심어라고 한다. 함께 등장한다는 동시출현(co-occurrence)은 두 키워드가 어떤 기준의 범위(문장, 문단, 소절, 문자열, 문헌전체 등) 내에서 함께 등장하는 것을 의미한다. 이 두 가지 개념은 언어 네트워크 분석에서 매우 중요하다. 즉, 언어 텍스트에서 의미를 나타내는 키워드들을 어떻게 선정할 것인가? 하는 문제(키워드 선정)와 두 단어가 동시출현하는 관계는 어떤 기준으로 파악할 것인가? 하는 문제(동시출현관계 파악)에 대한 적절한 기법을 적용하여야 한다는 것이다. 전자는 언어 네트워크에서 노드를 선정하는 문제이며, 후자는 링크를 선정하는 문제이기 때문이다.

부연하면, 분석대상의 언어 텍스트 집합에서 키워드들을 선정하고, 선정된 키워드들의 동시출현관계를 파악하는 것이 중요하다. 그렇게 선정된 키워드(노드)와 키워드들 간의 연결관계(링크)로 구성되는 것이 언어 네트워크이며, 기존의

사회 네트워크 분석(social network analysis; SNA)을 위해 개발된 각종 분석기법들을 활용하여 언어 텍스트의 구체적인 특성들을 분석하게 된다. 그동안 이러한 언어 네트워크 분석은 의미 네트워크 분석이라는 표현 이외에도 단어 네트워크 분석, 키워드 네트워크 분석, 언어 연결망 분석, 의미망 분석, 개념 연결망 분석, 네트워크 텍스트 분석, 동시단어 분석 등의 이름으로 다양한 영역에서 활용되어 왔다.

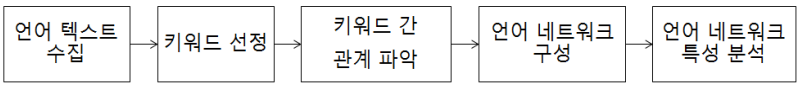
2.3 언어 네트워크 분석의 유용성

언어 네트워크 분석은 언어 텍스트의 내용에 포함된 단어들을 그대로 분석에 활용하는 정성적 분석이나 내용분석에서 발생하는 코딩 오류를 예방함과 동시에 사전에 정의된 단어들에 대한 빈도분석에 머무르는 언어분석의 한계점도 극복이 가능하다. 또한 메시지를 구성하고 있는 단어와 그들이 연결되는 패턴의 분석을 통해 메시지가 가지고 있는 본연의 내용 구조를 도출할 수 있으며, 이 과정에서 연구자의 주관적 오류를 어느 정도 배제할 수 있다는 장점이 있다. 아울러 언어 네트워크 분석은 단어와 단어를 연결하는 관계를 통해 네트워크를 형성하므로 단어 간 관계에 따른 개별 단어들의 상대적 위치를 정량적인 네트워크 지표로 산출할 수 있다. 이러한 단어 네트워크 분석의 유용성을 정리하면 다음과 같다(박치성, 정지원, 2013).

첫째, 언어 네트워크 분석의 가장 큰 장점은 텍스트를 해체한 후, 이를 다시 조합하여, 텍스트가 전달하고자 하는 행간의 의미를 파악함으로써, 그 텍스트가 전달하고자 하지만 명백히 드러나지 않는 주요 의미를 파악하는데 유용한 방법이다. 둘째, 네트워크 분석을 적용하는 경우 가장 큰 장점은 언어구조를 공간적으로 표시함에 따라, 텍스트에 나타난 주요개념과 다른 개념들과의 관계를 시각적으로 파악할 수 있다. 셋째, 네트워크 분석을 통하여 여러 종류의 네트워크 중심성(network centrality)이 높게 나타난 개념을 찾아냄으로써, 전체 텍스트가 전달하고자 하는 의도 및 의미를 이해할 수 있다. 넷째, 네트워크 분석을 통하여 단순히 특정 개념이 얼마나 많이 등장하였는지에 그치지 않고, 그 개념이 다른 개념들과의 관계에서 어떤 역할을 하는지, 또는 단어들이 특정패턴으로 배열되어 있는가에 대한 구조적 분석을 통하여, 특정한 의미의 순환구조, 특정 개념들 간의 공동의미 구성 등을 파악할 수 있다. 다섯째, 질적방법과 양적방법을 동시에 적용할 수 있다는 장점이 있다.

2.4 언어 네트워크 분석의 과정

언어 텍스트에 나타난 다양한 의미들을 나타내는 단어(키워드)는 언어 네트워크의 노드(node)가 되고, 단어(키워드)들 간의 관계는 언어 네트워크의 링크(link)가 된다. 이렇게 구성된 언어 네트워크는 사회 네트워크 분석에서 개발된 각종 분석기법을 적용하여 분석되고 해석이 된다. 언어 네트워크 분석은 언어 텍스트에서 선정한 키워드들을 이용하여 언어 네트워크를 구성하는 과정과 언어 네트워크의 특성을 분석하고 해석하는 과정으로 크게 구분된다. 이러한 전체적인 분석 과정을 도식화하면 〈그림 1〉과 같다.



〈그림 1〉 언어 네트워크 분석 과정

2.4.1 언어 텍스트 수집

언어 텍스트는 분석의 목적에 따라 다양한 유형의 텍스트가 수집된다. 분석의 목적에 따라 수집되는 텍스트의 주요한 유형을 구분하면 다음과 같다. 첫째, 문헌에 나타난 서지적 특성을 분석하는 경우 주로 문헌 텍스트(학술논문, 특허, 보고서 등)를 대상으로 한다. 둘째, 사회적 특성을 분석하는 경우 주로 사회적 메시지의 텍스트(언론기사, 구술자료, 미디어 텍스트 등)를 대상으로 한다. 셋째, 언어적 특성을 분석하는 경우 주로 특정 언어 주제를 나타내는 코퍼스 형태의 텍스트를 대상으로 한다. 넷째, 심리적 특성을 분석하는 경우 주로 심리적 특성 분석이 가능한 다양한 유형의 텍스트(연설문, 구술자료, 커뮤니케이션 자료 등)를 대상으로 한다.

2.4.2 키워드 선정

분석대상인 언어 텍스트로부터 키워드를 선정하는 작업에는 많은 세부적인 과정이 요구된다. 첫 번째 과정은 언어 텍스트로부터 단어들을 추출하는 작업(단어의 추출)이다. 단어는 언어 텍스트의 본문에서 추출하거나 언어 텍스트의 특정한 위치(제목, 목차, 초록, 주제어, 본문 등)에서 추출하게 된다. 그리고 단어의 추출은 분석목적에 따라 명사형(명사, 명사구) 단어, 동사/형용사형 단어들이 추출된다. 이렇게 추출된 단어들은 각종 정제 작업을 통해 보완하는 작업(단어의 정제)을 수행한다. 단어의 정제는 단어의 교정작업, 통제작업, 그리고 제거작업으로 구분한다. 단어의 교정작업은 추출된 단어에서 단/복수, 약어, 띄어쓰기, 품사형태 변경 등과 같은 클렌징(cleansing) 작업을 의미한다. 단어의 통제작업은 추출된 단어가 통제된 용어가 아닐 경우, 시소러스와 같은 통제어 사전을 이용하여 동의어, 유사어, 광의어, 협의어 등을 통제하여 적절한 단어를 채택하는 것을 의미한다. 시소러스와 같은 통제어 사전이 없는 경우, 해당 분야의 주제전문가로부터 도움을 받아가면서 단어를 통제할 수 있다. 그리고 단어의 제거작업은 출현빈도가 높은 단어들 중에서 너무 일반적인 개념을 나타내는 단어 등과 같이 의미있는 단어가 되지 못하는 단어들을 제거하는 작업을 의미한다.

두 번째 과정은 분석에 사용될 키워드를 선정하는 작업(키워드 선정)이다. 앞의 단계인 단어의 추출 과정은 키워드 선정의 전처리 과정이 된다. 언어 텍스트에서 키워드는 텍스트의 주제적 특성을 가장 잘 나타낸다고 판단되는 단어이기에, 대체로 언어 텍스트로부터 단어를 추출하고 그 단어들 중에서도 키워드로서 자질(features)이 높은 단어들로 선정하게 된다. 텍스트를 구성하는 모든 단어가 키워드가 되지 않고, 키워드로서 자질이 있는 단어들만 선택한다는 것이다. 키워드로서의 자질은 주로 해당 텍스트의 주제적 특성을 나타내는 것이므로, 단어의 출현빈도를 기초로 판단하는 방법을 많이 사용한다. 즉 고빈도 단어, 중빈도 단

어, 가중치(예: TF*IDF)가 높은 단어 등으로 키워드로서의 자질을 판단한다는 것이다. 이외에도 특정한 의미를 나타내는 단어들만 선정하는 방법, 비교대상의 그룹이 있을 경우 각 그룹별로 공통적으로 등장하는 단어들만을 키워드로 선정하는 방법 등도 가능하다. 이러한 특성도 키워드로서 자질이 될 수 있다는 의미이다. 이처럼 적절한 자질을 갖춘 단어들을 키워드로 선정하는 방법이 가장 일반적이라 할 수 있다.

한편, 언어 텍스트로부터 추출한 단어들로부터 키워드를 선정하기보다, 특정 주제의 범주체계에 해당되는 키워드를 언어 텍스트에 코딩(coding)하여 부여하는 방법이 있다. 특정한 주제범주체계는 대분류, 중분류, 소분류 등과 같은 프레임(frame) 구조를 가지며, 각 프레임은 특정한 개념을 나타내는 키워드 세트로 구성된다. 이러한 주제범주체계는 명확한 근거에 의해 결정하여야 한다. 주제범주체계를 분석대상의 언어 텍스트에 근거하지 않고 구성할 수 있으며, 결과에 따라 부여되는 키워드가 달라지기 때문이다. 주제범주체계가 결정되면, 언어 텍스트에서 단어들을 추출하여, 그들을 각 프레임에 해당되는 키워드로 코딩하는 작업을 해야 한다. 예를 들어, 문헌정보학이라는 주제영역을 대분류, 중분류, 소분류의 프레임 구조로 구분하고 각 프레임에 특정한 키워드를 선정한다. 그리고 분석대상의 언어 텍스트에서 단어를 추출하고, 이것을 주제범주체계의 해당되는 키워드로 코딩하게 된다. 또 다른 사례로, 언어 텍스트에 나타난 심리적 변인들의 관계를 분석할 경우, 심리적 변인을 나타내는 일반적인 주제범주체계를 미리 결정하고, 언어 텍스트에서 추출한 단어들을 해당되는 주제범주체계로 분류하는 경우도 여기에 해당된다. 만약 심리적 개념을 나타내는 '욕망'이라는 프레임이 있다면, 추출된 단어들 중에서 '욕심', '갈망', '욕구', '열망' 등의 개념을 나타내는 단어들은 모두 '욕망'이라는 프레임(또는 키워드)로 코딩한다는 것이다.

2.4.3 키워드 간 관계 파악

언어 네트워크 분석에서 키워드 간 관계는 대체로 동시출현관계로 파악이 된다. 동시출현(co-occurrence)은 공출현이라 하며, 어떤 대상이 특정한 기준의 범위 내에서 동시에 출현하는 것을 말한다. 따라서 키워드의 동시출현은 주어진 범위(문장, 텍스트 전체 등)의 언어 텍스트 내에서 키워드들이 동시에 출현하는 것이다. 이 때 하나의 텍스트 범위 내에서 동시에 등장하는 키워드들을 동시출현 키워드들이라 한다. 달리 설명하면, n개의 텍스트 집합($T_1 \sim T_n$)에 출현하는 m개의 키워드 집합($k_1 \sim k_m$)이 있을 경우, 두 키워드 k_i와 k_j가 특정한 텍스트의 주어진 범위 내에서 동시에 출현하는 경우, 키워드 쌍(k_i, k_j)은 동시출현관계에 있다고 한다. 동시출현관계의 단어를 공기어라고 한다. 그리고 키워드 쌍(k_i, k_j)이 동시출현하는 텍스트의 수(C_{ij})를 동시출현빈도라 한다. 언어 네트워크에서 키워드 쌍(k_i, k_j)은 링크로 연결될 수 있으며, 이들의 동시출현빈도는 두 키워드 노드의 연결강도(strength)로 표시된다.

키워드의 동시출현관계를 판단하는 텍스트의 범위는 주로 문자열(제목, 주제어 리스트 등), 문장, 문단, 소절, 또는 텍스트 전체(논문, 신문기사, 인터뷰나 발언 내용 전체 등) 등으로 다양하게 결정할 수 있다. 동시출현하는 키워드

들은 해당되는 범위의 텍스트 내에서 특정한 의미(개념)을 나타내는 주제를 표현하기 위해서 함께 사용되는 것이기에, 주제적으로 유사한 의미(개념)라고 판단하는 것을 전제로 한다. 즉, 문장을 범위로 할 경우, 한 문장 안에서 둘 이상의 키워드들이 함께 사용되고 있다는 것은 이들 키워드들이 해당 문장 내에서 주제적으로 서로 밀접한 관계를 가진다는 것을 의미한다. 그러므로 동시출현빈도에 의한 네트워크 관계의 설정은 충분히 타당성을 가진다(강명구, 2000). 키워드들의 동시출현관계와 빈도의 파악은 수작업으로 쉽지 않은 일이기에, 관련된 처리도구를 사용하여야 한다.

2.4.4 언어 네트워크 구성

키워드의 동시출현관계가 파악되면, 이를 토대로 키워드 동시출현빈도 행렬을 만들고, 이 행렬로 언어 네트워크를 구성할 수 있다. 언어 네트워크는 키워드의 동시출현관계로부터 구성된다는 것이다. 키워드 동시출현관계의 언어 네트워크를 구성하는 구체적인 절차는 다음과 같다. 첫째, 특정 범위의 텍스트별로 동시에 출현하는 키워드 리스트를 만든다. 둘째, 텍스트별로 등장하는 키워드 리스트는 '텍스트 × 키워드' 형태의 이원모드(2-mode) 리스트이므로, 이것을 '키워드 × 키워드' 형태의 일원모드(1-mode) 키워드 행렬로 변환한다. 이것을 키워드 동시출현 행렬(keyword co-occurrence matrix)이라고 한다.

키워드 동시출현 행렬에서 키워드 쌍(k_i, k_j)에 해당되는 셀의 값은 동시출현빈도를 나타내므로, 기본적으로 가중행렬의 형태가 된다. 분석의 목적에 따라 가중행렬에서 가중 네트워크(valued network)를 구성할 수 있고, 가중행렬을 이진행렬로 변환하여 만든 이진 네트워크(binary network)를 구성할 수도 있다. 가중행렬을 이진행렬로 변환하는 방법은 단순하게 특정한 기준값(cut-off point)에 따라 변환하거나, 유사도 계수를 이용하여 변환할 수도 있다. 후자의 경우, 키워드 간의 통계적인 연관성을 잘 평가할 수 있는 다른 유사도 계수(자카드 계수, 코사인 계수, 상관계수 등)를 적용한 다음 그 결과의 값에서 적절한 기준값을 적용하여 이진행렬로 변환하게 된다. 이러한 방식으로 언어 네트워크를 구성하게 되는데, 대체로 네트워크 분석도구에서 관련된 작업을 수행하는 기능을 사용하게 된다.

2.4.5 언어 네트워크 특성 분석

구성된 언어 네트워크(가중 네트워크 또는 이진 네트워크)를 대상으로 네트워크의 다양한 특성들을 분석한다. 언어 네트워크의 특성들은 네트워크 분석도구를 활용하여 분석하게 되는데, 주로 시각화 분석과 분석지표에 의한 분석을 한다. 네트워크 분석도구는 UCINET와 NetMiner, Pajek, NodeXL, Gephi, R의 SNA 패키지 등을 사용할 수 있다. 시각화 분석은 언어 네트워크를 시각화한 후, 시각적으로 나타나는 특성을 설명하는 것을 말한다. 네트워크의 시각화는 네트워크의 분석도구에서 제공하는 시각화 기능을 사용할 수 있다. 분석지표에 의한 분석은 다양한 유형의 분석지표를 사용하여 네트워크의 특성을 파악하는 것이다. 분석지표의 주요 유형은 기본속성 분석(밀도, 지름 등), 중심성 분석(연결정도 중심성, 근접 중심성, 매개 중심성 등), 하위 네트워크 분석(클러스터링 분석, 파당 분석, 구조적 등위성 분석 등),

에고 네트워크 분석 등으로 구분할 수 있다(이수상, 2012).

3. 학술논문의 내용분석

3.1 분석 대상과 범주의 설정

한국연구재단에서 제공하는 한국학술지인용색인 검색사이트인 KCI(www.kci.go.kr)를 기본으로 하여 '언어 네트워크 분석'이라는 키워드로 학술논문들을 검색하고, DBPIA, KISS, RISS와 같은 국내 학술논문 데이터베이스에서 검색결과의 원문을 확인하며, 내용을 검토하면서 네트워크 분석과 연관이 없는 논문들을 제외하고, 각 논문의 참고문헌에서 언어 네트워크 분석 관련 학술논문들을 추가하는 과정을 통해, 전체 53편의 분석대상의 학술논문을 선정하였다. 특정한 언어 텍스트로부터 언어 네트워크를 구성하고, 네트워크 분석도구와 지표를 사용하는 방법을 적용한 국내학술논문은 2007년부터 등장하기 시작하였고, 매년 꾸준히 증가하고 있다. 단독연구(약 30%)보다 공동연구(약 70%)의 비율이 훨씬 더 높다. 주제분야는 행정학/정책학 분야가 가장 많으며, 인문학에서부터 공학에 이르는 주제분야를 나타내고 있다. 연도별, 저자수별, 주제분야별 현황은 〈표 1〉에서 〈표 3〉에 정리하였다. 그리고 전체 38종의 학술지에 투고되었으며, 이 중에서 2편 이상 논문을 수록한 9개 학술지의 현황은 〈표 4〉와 같다.

〈표 1〉 연도별 구분

년도	2007	2009	2010	2011	2012	2013
편수	1	2	3	15	12	20

〈표 2〉 저자수별 구분

저자수	1	2	3	4	5
편수	16	21	12	3	1

〈표 3〉 주제분야별 구분

주제분야	편수	주제분야	편수
간호학	1	산업공학	1
경영학	3	신문방송학	4
경제학	1	심리과학	2
행정학/정책학	13	언어학	1
교육학 등	6	정치이론/정치외교학	2
기술정책	1	지구과학	3
문헌정보학/기록보존	5	지역학	2
사회과학일반	1	학제간연구	7

〈표 4〉 학회지별 구분(2편 이상 게재학회지)

학술지	편수	학술지	편수
지능정보연구	2	한국위기관리논집	2
한국도서관정보학회지	2	한국지구과학회지	3
한국도서연구	2	한국콘텐츠학회지	6
한국비블리아학회지	2	한국행정학보	3
한국심리학회지	2		

※ 1편 수록 학술지는 총 29편

53편의 학술논문들을 대상으로 언어 네트워크 분석 방법의 적용 실태를 파악하기 위해 다음과 같은 6가지 분석 범주를 사용하였다. 이 범주들은 〈그림 1〉의 언어 네트워크 분석 과정에서 요구되는 주요한 의사결정 사항들을 중심으로 구성하였다.

- 언어 텍스트의 유형
- 키워드 선정 방법
- 키워드 관계파악 방법
- 네트워크 구성 방법
- 네트워크 분석도구의 유형
- 네트워크 분석지표의 유형

3.2 언어 텍스트의 유형

언어 네트워크 분석에서 어떤 유형의 언어 텍스트를 분석대상으로 삼을 것인가는 매우 중요한 의사결정 사안이다. 분석의 목적이나 대상에 따라 수집되는 언어 텍스트의 유형이 다를 수밖에 없기 때문이다. 그리고 특정한 유형이 결정되고 난 다음, 그것을 어떻게 수집하고, 얼마만큼의 크기를 갖게 하는가도 결정하여야 한다. 분석대상의 학술논문들에서 살펴본 언어 텍스트의 유형은 〈표 5〉와 같다.

언어 텍스트의 유형으로 가장 많이 사용된 것은 학술논문으로 나타났다. 학술논문은 특정한 학술지(1종 이상)를 대상으로 하거나, 특정한 기간 동안 투고된 논문들을 대상으로 한다. 그리고 특정한 검색사이트를 통해 원하는 주제분야의 학술논문을 검색하여 사용하는 경우도 있다. 국외 검색사이트는 Web of Science, PubMed, ClinicalTrials.org 등이 활용되고 있으며, 국내 검색사이트는 RISS나 DBPIA, 녹색기술정보포털, 국회전자도서관 등이 활용되고 있다.

〈표 5〉 언어 텍스트의 유형

유형	건수	유형	건수
학술논문	18	석박사논문	1
인터뷰 자료	13	법령자료	1
신문기사	8	신약성경 본문	1
기록자료	5	연구대회 수상작품	1
혼합자료	2	과학교육과정/교과서의 특정내용(학습목표)	1
잡지기사	1	검색어 세트	1

두 번째 많이 사용한 유형은 인터뷰 자료이다. 특정한 기준에 의해 선정된 대상자들과 심층인터뷰로 얻어진 정성자료를 전사하고 문서화 작업을 수행하여 텍스트로 구성하고 있다. 인터뷰 대상 집단은 특정한 표적집단이 선택되기도 하지만, 필요시 분석하고자 하는 사안과 연관이 있는 이해관계자들로 선택되고 있다. 인터뷰 자료는 주로 직접 면담으로 수집하지만, 필요시 개방형 설문지를 통해 수집하고 있다.

그 다음으로는 신문기사, 기록자료, 토론자료의 순으로 조사되었다. 신문기사의 경우, 주로 특정한 주제어로 검색사이트(카인즈, 네이버, 다음 등)에서 검색하여 사용하고 있다. 검색된 결과는 전체 기사를 사용하거나 특정한 기사만(논설, 사설, 칼럼 등)을 사용하고 있다. 신문기사는 특정한 미디어(한겨레와 조선일보)에 한정하거나 특별히 한정하지 않고 있다. 기록자료는 특정 기관(초등학교)에서 관할 기관(지역교육청) 사이에 접수하고 발송한 공문서 전체를 대상으로 하거나, TV토론 참가자의 발언자료, 한 개인(대통령)의 재임기간 중의 발언록, 학회장 인사말, 연설문 등을 사용하고 있다. 토론자료는 특정한 주제(과학적인 것이란 무엇인가?)에 대하여 복수의 인원이 집단으로 토론하여 그 내용을 텍스트로 전사한 것, 선거 후보자가 특정한 주제(통일외교)에 대해 합동토론한 내용을 전사한 것들이 여기에 포함된다.

그리고 2건 이상의 혼합자료, 잡지기사, 석박사논문, 법령자료, 신약성경 본문, 연구대회 수상작품, 과학교육과정/교과서의 특정내용(학습목표), 검색시스템의 검색어 세트와 같이 다양한 유형의 언어 텍스트가 사용되고 있다.

3.3 키워드 선정 방법

언어 텍스트가 수집되고 나면, 키워드를 선정하게 된다. 2장의 키워드 선정과정에서 살펴보았듯이, 키워드 선정에는 여러 가지 고려사항에 대한 파악이 요구된다. 대부분의 분석대상 논문들은 언어 텍스트에서 단어를 추출하고, 주제적 자질이 높은 단어를 키워드로 선정하는 방법을 사용하고 있다. 이 경우의 고려사항은 다음과 같다: 언어 텍스트의 어떤 요소로부터 단어를 추출할 것인가? 어떤 품사의 단어를 선정할 것인가? 단어 추출에 사용되는 도구(프로그램)는 무엇인가? 단어의 정제작업은 어떻게 수행하는가? 키워드 선정에 사용된 자질은 무엇인가? 한편, 언어 텍스트로부터 직접 추출한 단어들에서 키워드를 선정하는 방법이 아니라 특정 주제의 범주체계를 별도로 구성하고, 언어 텍스트에서 추출한 단어를 해당 주제범주체계의 키워드로 코딩하는 방법을 사용하는 경우에서 고려사항은 주제범주체계를 구성하는 방법과 키워드의 코딩 방법을 파악하였다.

분석대상의 학술논문들에서 이러한 고려사항에 대해 상세하게 언급하지 않은 경우가 대부분이었다. 일부라도 관련된 언급이 있는 경우, 그 구체적인 내용을 파악한 결과는 〈표 6〉과 같다.

단어 추출은 언어 네트워크 분석에서 매우 중요한 과정의 하나이다. 단어는 언어 텍스트에서 의미있는 개념을 나타내는 것이기에, 이것을 추출하는데 신중을 기해야 할 것이다. 어떤 단어를 추출할 것인가에 따라 향후 분석의 결과에 많은 영향을 미치기 때문이다. 가장 많이 나타난 단어의 추출 대상 요소는 텍스트의

〈표 6〉 키워드 선정 방법

고려사항	내용
단어 추출 대상 요소	본문(21건), 저자부여 키워드(5건), 제목(5건), 기타(3건), 검색질의(1건)
추출되는 단어의 품사	명사(5건), 고유명사 제외한 단어(1건), 단어 또는 구절(1건)
단어 추출 프로그램	KrKwic(30건), 지능형형태소분석기(3건), 글잡이(1건), CiteSpace(1건)
단어 정제작업 사용	12건
키워드 선정에 사용된 자질	상위 출현빈도(12건), 특정 빈도수 이상(10건), 문헌빈도(1건), 공출현빈도(1건), 특정 주제의 키워드 선정(6건)
특정 주제의 범주체계 사용	주제범주체계 사용(3건), 디스크립터 세트 사용(1건)

본문(21건)이었다. 구체적으로 살펴보면, 인터뷰 자료의 본문에서 추출하는 경우가 가장 많았고(10건), 그 다음은 신문기사의 본문이었다(4건). 이외에도 성경 텍스트의 본문, 합동토론회자료/보고서/인사말/연설문/법령 등의 본문이 대상이 되고 있었다. 특별히 언급하지 않은 논문들도 제법 나타났다. 추출된 단어의 품사는 명사 또는 명사구라고 언급하는 경우가 그 중에서는 많은 편이며, 단어 추출 프로그램으로는 KrKwic를 가장 많이 사용하고 있다. 그리고 단어의 정제작업을 수행하였다고 조금이라도 언급하는 경우는 12건이다.

키워드 선정 기준으로 사용되는 단어의 자질 요소는 주로 단어의 출현빈도를 사용하고 있으며, 상위 순위의 단어, 특정 빈도수 이상의 자질을 가지는 단어를 키워드로 선정하고 있다. 그리고 특정한 주제의 범주에 해당되는 키워드만 선정하는 경우(6건)도 있다. 후자의 경우는 '온라인 쇼핑 형태와 관련된 명사'(이동일, 이혜준, 2012), '통일외교정책을 구성하는 단어'(박성희, 2009), '윤리관련 주요 단어'(김만재, 전방욱, 2012) 등이 사례가 된다. 이 경우 키워드는 복수의 집단을 구분하여, 집단별로 선정하고, 복수의 네트워크를 구성하기도 한다. 남자집단의 키워드와 여자집단의 키워드, 특정하게 구분된 분야별로 키워드 선정, 질문 영역별 키워드 선정 등과 같은 사례가 해당된다. 이러한 집단 각각은 추후 네트워크를 구성할 때 별도의 네트워크가 된다.

한편, 주제범주체계를 사용한 경우는 4건인데, 이 중에서 3건은 특정한 주제의 범주체계(프레임 구조)를 사용하는 경우이고, 나머지 1건은 특정한 주제의 디스크립터 세트를 사용하는 경우이다. 첫 번째 사례는 원자력발전소 입지 갈등의 핵심 당사자인 지역주민의 인식 프레임의 특성을 파악하는 것으로 갈등 프레임과 이슈 프레임의 2가지 주제에 대한 프레임 7개와 2개, 그리고 각각 41개와 8개의 세부 프레임을 키워드로 선정하고 있다(심준섭, 김지수, 2011). 그런 다음 인터뷰 텍스트에 추출한 단어 또는 구절을 각 세부 프레임별로 분류작업을 하는 코딩 작업을 수행하고 있다. 두 번째 사례는 2009 개정 과학교육과정의 지구과학 I 목표와 관련 교과서의 학습목표와의 일치성을 알아보기 위한 것으로, 국내외의 문헌 연구 및 전문가 자문을 통하여 분석 프레임을 3가지 대범주(능력, 공통개념, 행위동사)에 28개의 하위범주를 설정하고, 과학교육과정과 2개의 교과서에 나타난 목표기술 내용을 대상으로 코딩작업을 수행하고 있다(정덕호 외, 2013). 세 번째 사례는 대

학 성인학습자 12명의 면담내용을 내용분석하는 것으로, 인터뷰 내용을 수차례 반복하여 읽으면서 세부항목에 대한 코딩과 범주화 작업을 통해 61개의 개념을 도출하고, 이것을 키워드로 선정하는 경우이다(현영섭, 신은경, 2011). 그리고 디스크립터 세트를 사용한 사례는 '전자기록'이라는 연구영역에 해당하는 대분류-중분류-소분류의 디스크립터 세트로 주제범주체계로 구분하고, 이러한 디스크립터들을 분석대상인 학술논문들에 부여하며, 각 디스크립터에 해당하는 학술논문들에서 키워드들을 선정하여, 디스크립터-키워드의 프로파일을 만들어 분석하는 경우이다(김판준, 서혜란, 2012).

3.4 키워드 관계파악 방법

언어 텍스트에서 선정된 키워드들을 대상으로 관계를 파악하게 되는데, 이 작업은 관계의 부여 기준, 관계의 경계 범위, 관계의 강도, 관계의 차원, 관계의 표현과 같이 몇 가지로 나누어 살펴보았다.

첫째, 관계를 부여하는 기준은 대부분 동시출현빈도를 가장 많이 사용하고 있다. 유사용어로는 공출현빈도, 공출현, 동시출현관계, 단어동시출현 등을 사용하고도 있다. 그러나 대부분 동시출현빈도를 사용한다는 언급 정도에서 그치고 있다. 일부 논문에서는 동시출현빈도의 값을 대상으로 코사인 유사계수를 적용하고, 키워드 사이의 행렬을 구축하는 것과 같이 관계의 형성과정을 보다 상세히 설명하는 경우도 있다. 그리고 동시출현관계의 정도를 t-score 값으로 계산하여 특정 기준값 이상인 경우 관계를 부여하는 방법(김혜영, 이도길, 강범모, 2011), 설문조사에 의해 관계를 부여하는 방법(정은경, 정혜승, 손영우, 2011) 등도 있다.

둘째, 관계의 경계 범위는 동시출현이라는 관계가 형성되는 범위(경계)를 문장 단위라고 언급한 경우 이외에는 대부분 특별한 언급을 하지 않고 있다. 학술논문인 경우 대체로 저자가 부여한 키워드를 사용하므로, 학술논문 자체가 관계의 범위가 된다.

셋째, 관계의 강도는 동시출현빈도 2회 이상과 같이 특별히 가중관계를 사용하는 경우에 언급하고 있다. 넷째, 관계의 차원은 대부분 일원모드이지만, 일부는 이원모드의 관계에서 일원모드의 관계로 변환하여 사용하고 있다. 일원모드로 변환하기 위해 코사인 유사계수와 같은 유사도 계수를 적용하고 있다. 다섯째, 관계의 표현을 위해서는 대부분 행렬을 사용하고 있다. 키워드 간의 동시출현빈도가 행렬 셀의 값이 되는데, 이러한 행렬을 만드는 도구로 KrTitle 프로그램을 사용한다고 언급하는 경우가 일부 있다.

3.5 네트워크 구성 방법

키워드 간의 관계의 행렬이 만들어지면, 이를 토대로 언어 네트워크를 구성하게 된다. 언어 네트워크는 키워드를 노드로 하는 이진 네트워크 또는 가중 네트워크 형태로 구성된다. 이진 네트워크는 특정한 동시출현빈도 이상 또는 유사도 계수에서의 기준값 이상을 가지는 키워드들을 연결하는 경우이다. 또한 이원모드 행렬에서 변환하여 언어 네트워크로 구성할 수 있는데, 이 경우의 네트워크를 준연결 네트워크(quasi-network)라 한다. 이원모드에서 일원모드로의 변환을 위해서는 코사인 유사계수

〈표 7〉 네트워크의 종류

네트워크의 집단	단수의 네트워크(22건) 복수의 네트워크(31건) 이원모드의 네트워크(5건)
네트워크의 강도	가중 네트워크(5건)
네트워크 노드의 크기	최소 8개에서 최대 4,521개

와 같은 적절한 유사도 계수를 적용하여야 한다. 동시출현관계의 값을 그대로 사용하여 관계의 강도를 표시할 경우 가중 네트워크가 된다. 각 네트워크는 선정된 키워드의 수만큼 노드의 크기를 가진다.

〈표 7〉에서 알 수 있듯이, 각 논문에서 단수의 네트워크를 구성하는 것보다 복수의 네트워크를 구성하여 비교하는 방식으로 논의를 전개시키는 경우가 더 많다. 특히 복수의 네트워크들을 구성하는 경우 대체로 2개 이상의 네트워크들을 구성하고 있으며, 최대 8개의 네트워크를 구성하는 경우도 있다. 이원모드 네트워크를 구성한 경우는 전체 5건으로, 이원모드의 행렬에서 적절한 유사도 계수를 적용하여 일원모드의 행렬로 변환하는 과정을 통해 일원모드 네트워크로 변환하고 있다.

대부분의 네트워크는 이진 네트워크 형태이며, 가중 네트워크를 사용한 경우는 5건이 조사되었다. 가중 네트워크의 분석은 시각화 분석에 머물고 있는 것과 가중 네트워크를 대상으로 특화된 분석을 하는 경우가 혼재하고 있다. 그리고 각 논문에서 네트워크의 수가 최소 1개에서 8개이므로, 각 네트워크의 노드수를 파악하고 비교하는 것이 의미가 없을 수 있다. 그러나 언어 네트워크 분석에서 분석대상의 노드수가 대략 몇 개 정도인지 알고자 하는 수준에서 보면, 최소 8개에서 최대 4,521개에 이르고 있음을 알 수 있다. 20개 이상에서 50개 수준의 네트워크가 가장 많았다. 그리고 100개 이상인 경우는 드물게 나타났다. 그리고 특별히 노드의 개수를 언급하지 않은 경우도 제법 많았다.

3.6 네트워크 분석도구의 유형

각 논문들에서 저자들은 다양한 형식으로 사용된 네트워크 분석도구를 제시하고 있다. 따라서 약간의 용어통제를 통해 분석도구의 사용현황을 정리하면 〈표 8〉과 같다. UCINET에

〈표 8〉 네트워크 분석도구의 사용현황

구분	횟수	구분	횟수
NetMiner	19	UCINET/Pajek	3
UCINET/NetDraw	15	UCINET/NodeXL	2
NodeXL	4	UCINET/NetMiner	1
Pajek	4	UCINET/Gephi	1
Gephi	1	UCINET/NetDraw/NodeXL	1
R의 SNA 패키지	1	언급없음	1

는 NetDraw를 포함하고 있으므로, UCINET, UCINET/NetDraw, NetDraw라고 언급한 경우는 모두 UCINET/NetDraw을 사용한 것으로 간주하였다. 두 가지 이상의 분석도구를 언급한 경우(예를 들어 UCINET과 NodeXL을 함께 사용) 그대로 인정하였다. 이런 방식으로 통제할 경우, 가장 많이 사용된 분석도구는 NetMiner이며, 그 다음은 UCINET/NetDraw이다. 그리고 NodeXL과 Pajek도 제법 사용하는 편이었다.

3.7 네트워크 분석지표의 유형

각 논문들에서 저자들이 사용하였다고 제시한 네트워크 분석지표의 유형은 매우 다양하게 나타났다. 아주 단순하게 시각화 분석만 수행한 연구에서부터, 다양한 유형의 분석지표들을 활용하는 논문들도 있다. 이러한 유형을 정리하면 다음과 같다.

첫째, 네트워크 수준에서의 분석은 주로 밀도(density)의 분석에 머무르고 있다. 그리고 평균경로거리, 평균도달거리와 같은 거리(distance) 지표의 분석, GINI계수, 응집성, 분열성, 지름, 집중도(centralization) 등을 분석하는 경우도 있다.

둘째, 중심성 분석은 연결정도 중심성 지표를 사용하는 경우가 가장 많았다. 이처럼 한 가지 중심성 지표만을 분석하는 경우, 근접 중심성과 아이겐벡터 중심성의 분석도 있었다. 두 가지 중심성 지표를 분석하는 경우는 연결정도 중심성과 매개 중심성, 연결정도 중심성과 근접 중심성, 연결정도 중심성과 위세 중심성, 위세 중심성과 매개 중심성 등 다양한 형태를 나타내고 있었다. 그리고 세 가지의 중심성 지표를 분석하는 경우는 연결정도/근접/매개 중심성을 분석하는 경우, 연결정도/매개/페이지랭크 중심성을 분석하는 경우, 그리고 연결정도/매개/아이겐벡터 중심성을 분석하였다. 4가지 중심성(연결정도/근접/매개/아이겐벡터 중심성) 지표를 사용하는 경우도 있었다. 이처럼 중심성 분석지표의 사용 유형은 다양하며, 뚜렷한 기준이나 근거가 없어 보였다.

셋째, 하위 네트워크 분석은 클러스터 분석, 컴포넌트 분석, 파당 분석, 커뮤니티 분석, 구조적 등위성 분석, k-core 분석 등을 사용하고 있다. 경우에 따라서는 덴드로그램이나 다차원척도(MDS)와 같이 시각화된 군집분석을 보여주기도 하였다.

넷째, 아주 소수의 사례이지만, 에고 네트워크 분석을 시도한 경우도 있었다. 물론 에고 네트워크에 대한 시각화 분석 수준에서 머물거나 각 노드의 연결정도 정도만 나타내는 수준이었다. 다섯째, 가중치(weight)가 반영되는 연결정도 중심성, 표준화 연결정도 중심성, 전역중심성, 지역중심성, 매개중심성을 분석하는 경우도 있었다.

4. 결 론

본 논문에서는 국내 언어 네트워크 분석 분야 학술논문 53편을 대상으로 언어 네트워크 분석 방법의 적용실태를 내용분석을 통해 살펴보았다. 내용분석을 위한 범주항목은 언어 텍스트의 유형, 키워드 선정 방법, 키워드 관계파악 방법, 네트워크 구성 방법, 네트워크 분석도구의 유형, 네트워크 분석지표의 유형의 6가지로 선정하였

다. 각 범주항목별 주요 특성을 요약하면 다음과 같다.

언어 텍스트의 유형으로 학술논문을 가장 많이 사용하였으며, 그 다음은 인터뷰 자료 신문기사 등이다. 학술논문은 특정한 학술지(1종 이상)를 대상으로, 특정한 기간 동안 투고된 논문들을 대상으로 하거나, 특정한 검색사이트를 통해 학술논문을 검색하여 사용하고 있다. 인터뷰 자료는 특정한 기준에 의해 선정된 대상자들과 심층인터뷰로 얻어진 정성자료를 전사하여 문서화 작업을 수행하여 텍스트로 구성하고 있다. 그리고 신문기사, 기록자료, 토론자료 등을 사용하고 있다. 분석목적에 따라 언어 텍스트의 대상과 유형이 결정된다. 어떠한 유형의 언어 텍스트를 대상으로 하더라도, 적절한 탐색도구를 사용하여 수집하여야 할 것이다.

키워드 선정 방법에서, 텍스트의 본문에서 단어를 추출하고, 추출된 단어의 품사는 주로 명사이며, 단어 추출을 위해 KrKwic 프로그램을 가장 많이 사용하는 것으로 조사되었다. 추출된 단어는 정제작업을 거치며, 단어의 출현빈도라는 자질을 사용하여 키워드로 선정되는 경우가 많았다. 그리고 특정한 주제의 범주에 해당되는 키워드만 선정하는 경우도 나타났다. 한편, 분석대상의 주제를 나타내는 별도의 주제범주체계를 구성하여, 여기에 속한 키워드 세트를 사용하는 유형도 나타났다.

키워드 간 관계의 파악은 거의 동시출현빈도를 사용하고 있으며, 이를 통해 동시출현관계의 행렬을 만들고 있다. 두 키워드가 동시출현이라는 조건을 만족하는 범위(경계)는 분석목적에 따라 하나의 텍스트 전체를 사용하든지, 텍스트 내의 문장, 문단 등과 같은 특정한 범위를 설정하고 있다.

네트워크 구성 방법에서 보면, 단수의 네트워크보다 복수의 네트워크를 구성하고 있으며, 각 네트워크는 키워드의 동시출현관계의 행렬에서 직접 언어 네트워크를 구성하거나, 이원모드 행렬에서 적절한 유사도 계수를 적용하여 일원모드의 행렬로 변환하는 과정을 통해 일원모드 네트워크로 변환하여 구성하기도 하였다. 대부분은 이진 네트워크 형태이며, 가중 네트워크를 사용한 경우는 조사되었다.

네트워크 분석을 위해 가장 많이 사용된 도구는 NetMiner이며, 그 다음은 UCINET/NetDraw, NodeXL, Pajek 등을 사용하였다. 그리고 네트워크 분석지표는 네트워크 수준에서의 분석, 중심성 분석, 하위 네트워크 분석 등에 이르는 지표들을 사용하고 있다. 네트워크 수준에서의 분석은 주로 밀도(density)의 분석에 머무르고 있다. 중심성 분석지표의 사용 유형은 다양하지만, 주로 연결정도 중심성, 근접 중심성, 그리고 매개 중심성 지표를 많이 사용하고 있다. 하위 네트워크 분석은 클러스터 분석, 컴포넌트 분석, 파당 분석, 커뮤니티 분석, 구조적 등위성 분석, k-core 분석 등을 다양하게 사용하고 있다.

지금까지 연구된 언어 네트워크 분석 관련 논문들의 내용분석에서 파악한 특성들은 다음과 같이 정리할 수 있다. 첫째, 학술논문과 인터뷰 자료를 분석대상의 언어 텍스트로 많이 사용하고 있다. 그리고 신문기사, 기록자료, 토론자료 등도 사용되고 있다. 둘째, 언어 네트워크에서 노드가 되는 키워드는 텍스트의 본문에서 단어를 추출하고, 정제작업을 거친 후, 단어의 출현빈도와 같은 키워드 자질을 사용하여 선정되는 경우가 많았다. 그리고 특정한 주제의 범

주체계에 해당되는 키워드로 선정하는 경우도 가능하다. 셋째, 키워드 간 관계의 파악은 거의 동시출현빈도를 사용하고 있으며, 이를 통해 동시출현관계의 행렬을 만들고 있다. 넷째, 언어 네트워크는 단수의 네트워크보다 복수의 네트워크를 구성하고 있으며, 대부분은 이진 네트워크 형태이며, 가중 네트워크를 사용한 경우도 있다. 다섯째, 네트워크 분석을 위해 많이 사용된 도구는 NetMiner와 UCINET/NetDraw, NodeXL, Pajek 등 사회 네트워크 분석에 많이 사용하는 도구들이다. 여섯째, 네트워크 수준에서는 주로 밀도(density)의 분석, 중심성은 주로 연결정도/근접/매개 중심성의 지표의 분석, 다양한 유형의 하위 네트워크 분석지표 등 사회 네트워크 분석지표들을 사용하고 있다.

이러한 특성들은 언어 네트워크 분석 방법론의 기초적인 체계를 구성하는 중요한 개념들이 된다. 즉 언어 네트워크 분석 방법의 기초적 체계는 언어 네트워크의 분석대상이 되는 언어 텍스트의 다양한 유형의 파악, 언어 텍스트에서 두 가지 키워드 선정 방법의 이해, 키워드들 간의 동시출현관계의 부여, 단수 네트워크와 복수 네트워크 또는 이진 네트워크와 가중 네트워크의 형태로 언어 네트워크의 구성, 기존의 사회 네트워크 분석도구와 분석지표를 활용한 분석과 같은 영역을 포함한다는 것이다.

참 고 문 헌

강명구 (2000). 정치뉴스에 나타난 한국 정치권력구조의 네트워크 분석 - '동시출현빈도'의 타당성 검증. 언론정보연구, 37, 93-130.

김동렬 (2013). 의미 네트워크 분석법을 활용한 초등 예비교사들이 생각하는 과학에 대한 의미 분석. 초등과학교육, 32(3), 327-345.

김만재, 전방욱 (2012). 언어네트워크 분석 기법을 활용한 인간배아복제 신문보도 분석. 생명윤리, 13(2), 19-34.

김유호 (2012). 의료민영화 논의에 따른 이슈용어의 연결 중심성 분석. 한국콘텐츠학회논문지, 12(8), 207-214. http://dx.doi.org/10.5392/JKCA.2012.12.08.207

김판준, 서혜란 (2012). 프로파일링 기법을 이용한 국내 기록관리학 분야 지식구조 분석. 한국기록관리학회지, 12(2), 29-50.

김혜영, 이도길, 강범모 (2011). 사건명사의 공기어 네트워크 구성과 분석. 언어와 언어학, 50, 81-106.

박성희 (2009). 제17대 대통령 후보 합동 토론 언어네트워크 분석. 한국언론정보학보, 45(1), 220-254.

박치성, 정지원 (2013). 텍스트 네트워크 분석-사회적 인식 네트워크. 정부학연구, 19(2), 73-108.

심준섭 (2012). 제주 해군기지 건설을 둘러싼 지역주민과 공무원의 갈등 프레임 비교분석. 행정논총, 50(4), 221-249.

심준섭, 김지수 (2011). 원자력발전소 주변 지역주민의 갈등 프레임 분석: 후쿠시마 원전사고의 영향을 중심으로. 한국행정학보, 45(3), 173-203.

이동일, 이혜준 (2012). 소비자 집단 인터뷰에서 의미 네트워크 응집 구조의 이해. 소비자학연구, 23(2), 249-272.

이수상 (2012). 네트워크 분석 방법론. 서울: 논형.

이창환, 심정미, 윤애선 (2005). 언어적 특성을 이용한 '심리학적 한국어 글분석 프로그램(KLIWC)' 개발 과정에 대한 고찰. 인지과학, 16(2), 93-121.

이혜준, 이동일, 이주현 (2010). 의미 네트워크 분석을 통한 프랜차이즈 교육 프로그램 개발. 경영교육연구, 14(2), 105-128.

정덕호 외 (2013). 언어네트워크분석을 이용한 교육과정 목표와 교과서 학습 목표와의 일치성 분석- 2009 개정 교육과정의 지구과학 I을 중심으로 -. Jour. Korean Earth Science Society, 34(7), 711-726.

정석환 (2013). 국정최고의사결정자의 정책신념에 관한 연구(Ⅲ): 제16대 노무현대통령의 부동산정책을 중심으로. 한국콘텐츠학회논문지, 13(6), 202-211.
http://dx.doi.org/10.5392/JKCA.2013.13.06.202

현영섭, 신은경 (2011). 대학 성인학습자의 학습저해 요소와 학습저해 해소 방법에 대한 개념 연결망 분석. 한국HRD연구, 6(3), 23-48.

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Chung, Duk Ho, Lee, Jun-Ki, Kim, Seon Eun, & Park, Kyeong Jin (2013). An analysis on congruency between educational objectives of curriculum and learning objectives of textbooks using semantic network analysis - Focus on earth science I in the 2009 revised curriculum -. Journal of Korean Earth Science Society, 34(7), 711-726.

Hyun, Young-Sup, & Shin, Eun-Kyung (2011). The concept networks of adult learners' factors inhibiting learning and methods resolving the factors in higher education. Journal of Korean HRD Research, 6(3), 23-48.

Jung, Seok-Hwan (2013). Study on the government chief decision maker's policy belief (Ⅲ): Focusing on 16th president Noh Moo-hyun's real estate policy. The Journal of the Korea Contents Association, 13(6), 202-211.

Kang, Myungkoo (2000). A network analysis of political power structure. Journal of Communication Research, 37, 93-130.

Kim, Dong-Ryeul (2013). An analysis of scientific concepts pre-service elementary school teachers have through semantic network analysis. Journal of Korean Elementary Science Education, 32(3), 327-345.

Kim, Hye-Young, Lee, Do-Gil, & Kang, Beom-Mo (2011). A network of co-occurring nouns of event nouns. Language and Linguistics, 50, 81-106.

Kim, Manjae, & Jun, Bang-Ook (2012). Semantic network analysis of three major human embryo cloning cases. Journal of the Korea Bioethics Association, 13(2), 19-34.

Kim, Pan Jun, & Suh, Hye-Ran (2012). A study on the analysis of intellectual structure of electronic records research in Korea using profiling. Journal of Records Management & Archives Society of Korea, 12(2), 29-50.

Kim, You-Ho (2012). Analysis of connection centrality degree of hot terminologies according to the discourses of privatizaiton of health care. The Journal of the Korea Contents Association, 12(8), 209-214.

Lee, Chang H., Sim, Jung-Mi, & Yoon, Aesun (2005). The review about the development of Korean linguistic inquiry and word count. Korean Journal of Cognitive Science, 16(2), 93-121.

Lee, Dong Il, & Lee, Hyejun (2012). Understanding the semantic network structure in the consumer group interview with the subnetwork analysis. Journal of Consumer Studies, 23(2), 249-272.

Lee, Hyejun, Lee, Dong Il, & Lee, Juhyun (2010). Development of franchise education program through semantic network analysis. Korean Business Education Review, 14(2), 105-128.

Lee, Soosang (2012). Network analysis methods. Seoul: Nonhyoung.

Park, Chisung, & Jung, Jiwon (2013). Text network analysis: Detecting shared meaning through socio-cognitive networks of policy stakeholders. Journal of Governmental Studies, 19(2), 73-108.

Park, Sung-Hee (2009). Semantic network analysis of presidential debates in 2007 election in Korea. Korean Journal of Communication & Information, 45(1), 220-254.

Shim, Junseop (2012). Comparison of conflict frames between local residents and bureaucrats over construction of the Jeju Naval Base. Korean Journal of Public Administration, 50(4), 221-249.

Shim, Junseop, & Kim, Jisoo (2011). Understanding Conflict Frames about a Neclear Power Plant: Focusing on the Effect of the Fukushima Nuclear Accident. Korean Public Administration Review, 45(3), 173-203.