『뉴스 코어 시소러스』의 구축 및 활용 방안에 관한 연구

A Study on the Establishment and Applications of the "News Core Thesaurus"

장 인 호(Inho Chang)*

〈목 차〉

초 록

본 연구는 뉴스 정보의 효율적인 색인 작성과 검색을 위한 뉴스 코어 시소러스를 구축하고 활용 방안을 제시한다. 뉴스 코어 시소러스는 뉴스의 모든 주제를 커버할 수 있는 매크로시소러스로 구축하여, 향후 정치, 경제 사회, 문화 등의 마이크로시소러스를 부분집합으로 가질 수 있도록 하였다. 뉴스 코어 시소러스는 디스크립터 2,012어, 비디스크립터 74어를 SKOS(Simple Knowledge Organization System)로 구현하였다. 일간신문이 아닌 주간이나 격주간 등의 정보량이 적고, 특수한 주제를 다루는 신문은 특수 주제영역만 상세히 다루고, 대형의 뉴스 아카이브나 포털 사이트에서는 각각의 마이크로시소러스를 병합 또는 통합하여 활용할 수 있는 방안을 제시한다.

키워드: 뉴스 시소러스, 뉴스 코어 시소러스, 매크로시소러스, 마이크로시소러스, SKOS, 활용방안

ABSTRACT

This study suggests the establishment and applications of the News core thesaurus for efficient indexing and searching of news information. News core thesaurus was constructed as macrothesauri which can cover all of news subjects and then has microthesauri like politics, economy, society, culture, etc. as its subsets. In this research, News core thesaurus embodied 2,012 descriptors and 74 non-descriptors by SKOS(Simple Knowledge Organization System). It suggests measures that treat only special subjects in detail in weekly newspaper or biweekly newspaper with little information and special subjects, which is not daily newspaper, and use each microthesauri by merging or integrating in huge news archives or portal sites.

Keywords: News thesauri, News core thesaurus, Macrothesauri, Microthesauri, Simple knowledge organization system, Application methods

* 대진대학교 인문과학대학 문헌정보학과 강사(hoinchang@gmail.com)

• 논문접수: 2013년 8월 28일 • 최초심사: 2013년 9월 2일 • 게재확정: 2013년 9월 15일

Ⅰ. 서 론

1. 연구의 배경 및 목적

신문에는 삼라만상의 주제를 가지고 있다. 그렇다고 수십, 수백만 어휘를 시소러스로 구축해야 하는 것은 아니다. 시소러스의 어휘는 정보량에 따라 증가하지만, 어디까지나 통제된 주제어이기 때문에, 만년필, 붓, 샤프펜슬이 아니라 '필기도구'만의 어휘로 충분한 경우도 많다. 붓은 필기구의 종류임에는 틀림없지만, 시소러스에서 가치 있는 용어로 사용할 정도의 주제 분야는 매우 드물 것이다. 물론 문방구 시소러스를 구축한다면 매우 중요한 키워드가 될 것이다. 우리나라에 구축된 시소러스는 많지 않지만 외국 사례와 비교해서 대형인 경우가 많다. 그렇다고 정보량이 많아서도 아니다.

같은 영역이라고 하더라도 신문은 간행주기에 따라 일간, 주간, 격주간 등이 있기 때문에 뉴스의 정보량에 큰 차이가 난다. 또한, 특정 영역을 다루지 않고 모든 영역을 다루는 종합지, 경제 영역만을 다루는 경제지, 스포츠 영역을 다루는 스포츠 신문 등으로 나눌 수 있다. 종합지가 모든 영역을 다룬다고 해서 경제지보다 더 많은 용어를 가지고 있어야 하는 것은 아니다. 종합지에서 다루는 증권과 관련된 용어보다 경제지에서 다루는 증권 관련 용어의 상세도가 높아야 하기 때문에 정보량의 차이와 비례하여 용어는 비슷할 수도, 어떤 면에서는 더 많을 수도 있을 것이다. 예를 들면, 같은 해에 공표된 "신문기사 종합시소러스"는 10,705어이고,[1] "경제신문 시소러스"는 13,646어이다.[2]

대학의 중앙도서관에 법률 관련 도서가 당연히 비치되어 있고, 분류기호에 의해 배가되어 있을 것이다. 그러나 법학전문도서관에서의 주제보다는 넓은 경우가 많을 것이다. 예를 들면, "준정"에 관한 논문이 중앙도서관에서는 색인어가 "가족법"으로 가능할 수 있지만, 법학전문도서관에서는 가족법으로 탐색해서는 너무 많은 탐색이 이루어질 수도 있을 것이다. 중앙도서관에서는 그냥 준정을 입력하면 자연어로 쉽게 탐색되기 때문에 - 가령 10여 종의 관련 도서가 있다고 가정 - 구태여 시소러스나 분류표에서 관리하지 않아도 될 것이다.

신문의 경우도 마찬가지이다. 예를 들어, 종합지에 안경테에 관한 기사가 분명히 존재할 것이지만, 이 정도의 수준을 시소러스로 관리한다면 수십만 이상의 용어 관리는 필수적이 될 것이다. 안경 관련 신문만을 관리하고자 하여 대형 시소러스를 만들 필요는 없다. 안경신문에도 정치문제가 있고, 경제문제도 있을 것이다. 그렇다고 이들 관련 용어를 상세히 다룰 필요는 없을 것이다. 핵심 영역을 상세히 다루고 그 밖의 영역은 넓게 다루는 것은 시소러스의 기본이다. 매크로시소러스와 마이크로시소러스의 전체와 부분집합의 관계를 사용하여 핵심 영역은 마이크로시소러스를 활용하

1) 한국언론연구원, 신문기사 종합시소러스(서울 : 한국언론연구원, 1993), p.iii.
2) 한국경제신문사, 경제신문 시소러스(서울 : 한국경제신문사, 1993.), p.3.

고 넓은 주제는 매크로시소러스를 사용하는 것이 보다 효율적일 것이다.

본 연구에서는 특정 분야를 대상으로 하지 않는 뉴스 코어 시소러스를 구축하여, 매크로시소러스로 이용함으로써 정보량에 따라 확장하거나 특정 영역의 마이크로시소러스와 조합하여 활용하는 방안을 제안한다.

정보량이 많지 않은 경우에 거대한 시소러스를 유지하는 것은 비용 면에서 비효율적이다. 처음부터 용어 수를 많이 수록하는 것은 좋지 않으며, 그렇게 많이 늘어나는 것도 아니다. 일본 뉴스시소러스의 경우 초판이 1975년에 간행되었으나 현재(2004년)에 제4판으로 약14,000어를 가지고 있다.[3] 그러나 우리나라의 경우 1993년 신문기사 종합시소러스가 10,705어이고, 2008년의 뉴스ML 시소러스가 26,782어이다.[4]

또한, 거대한 시소러스는 소규모의 신문기사나 방송뉴스를 관리하는 경우는 효율성이 떨어진다. 필요한 모듈만을 선택하여 사용할 수 있도록 할 필요가 있다. 국내의 경우 종합지와 경제종합지를 중심으로 개발되고 있지만 특수 주제 신문의 경우는 전무하다. 저비용으로 효율적으로 관리하기 위해서는 뉴스 코어 시소러스를 바탕으로 해당 특수 주제 분야를 확장해서 사용하는 것이 보다 효율적으로 사료된다.

본 연구는 뉴스의 모든 영역을 커버할 수 있는 뉴스 코어 시소러스를 구축하고, 각 뉴스 영역의 부분집합을 이루는 마이크로시소러스를 확장하거나 병합 또는 통합하여 활용하는 방안을 제시한다.

2. 연구의 절차

뉴스 코어 시소러스를 구축하고 매크로시소러스와 마이크로시소러스로 활용하기 위한 연구를 수행함에 있어서 자료의 수집 및 분석, 용어의 수집 및 통제, 카테고리의 설정, 코어 시소러스의 작성 및 활용 방안을 제시하고 결론을 맺는다.

본 연구의 상세한 절차는 다음과 같다.

첫째, 자료의 수집은 일차적으로 시소러스 규격, 참고서, 매크로시소러스와 마이크로시소러스의 관련 논문 등을 수집하고, 뉴스 관련 분류표, 시소러스, 정기간행물 등록현황 등을 수집하였다.

둘째, 자료의 분석은 특히, Lancaster,[5] Aitchison et al.,[6] Soergel[7]의 문헌을 대상으로 하였다.

---

3) 廣木守雄, 服部信司, ニュースシソーラス(東京 : 日外アソシエーツ, 2004).

4) 서명이 바뀌었으나 동일 기관에서 동일 목적으로 유지되어온 시소러스이다(한국언론재단. 뉴스ML 시소러스(상)(하). 서울 : 한국언론재단, 2008., 한국언론연구원. 신문기사 종합시소러스. 서울 : 한국언론연구원, 1993.).

5) F. Wilfrid Lancaster, *Vocabulary Control for Information Retrieval.* 2nd ed. (Arlington : Information Resources Press, 1986).

6) J. Aitchison, A. Gilchrist, and D. Bawden, *Thesaurus Construction and Use: a Practical Manual.* 4th ed. (London : Aslib imi, 2000).

7) D. Soergel, *Indexing Languages and Thesauri: Construction and Maintenance.* (Los Angeles : Melville Publishing Company, 1974).

Soergel에서 코어 디스크립터를, Lancaster와 Aitchison et al.에서 매크로시소러스와 마이크로시소러스의 논의를 분석하였다. 또한, 매크로시소러스와 마이크로시소러스의 학위논문이나 잡지기사를 분석하였다.

셋째, 카테고리는 신문기사 분류표나 시소러스의 카테고리, 그리고 문화체육관광부 및 시도에 등록된 신문[8]과 포털 사이트 다음의 '카테고리-신문'[9]을 분석하여 새롭게 매크로시소러스와 마이크로시소러스의 구성 방안을 도출하였다.

넷째, 뉴스 코어 시소러스를 구축하기 위한 설계 원칙을 제시하였다.

다섯째, 구축은 용어의 수집, 용어의 통제, 용어의 구조화, SKOS 파일 생성 등의 절차에 의하여 구축하였다. 용어 수집은 각종 뉴스 관련 분류표 및 시소러스를 참조하여 망라성을 확보하였다.

여섯째, 뉴스 코어 시소러스의 단독 사용 외에 다양한 활용 방안을 모색하였다.

Ⅱ. 이론적 배경

1. 코어 시소러스

코어 시소러스라고 하는 말은 일반적으로 받아들여지는 정의가 아니다. 하나의 구축 사례로 명명한 것이다. 코어 시소러스라고 하는 용어는 Soergel[10]의 코어 디스크립터, 캐나다 정부의 코어 주제 시소러스(Government of Core Subject Thesaurus)[11] 그리고 온톨로지 영역에서 코어 온톨로지라고 하는 용어에서[12] 영감을 얻어 명명하였다.

Soergel은 다양한 파일 조직화의 연결에 사용될 색인언어(분류표)의 설계를 위한 실제적인 제안을 하였다. 그는 단일 또는 복합 개념들로 구성하는 코어 분류을 설립하였다. 각각의 개념들은 코어 디스크립터라고 불렀으며, 용어나 기호법과 같은 독립된 기호로 표현하였다. 후조합시스템은 코어 디스크립터만이 디스크립터로서 사용되며, 전조합시스템은 코어 디스크립터들의 결합으로 형성된다고 하였다.[13]

8) 문화체육관광부(정기간행물등록현황) 〈http://www.mcst.go.kr/web/s_data/deptData/deptDataView.jsp?pSeq=81〉 [인용 2013. 5.3].
9) Daum 디렉토리 홈페이지 〈http://directory.search.daum.net/site_list.daum〉 [인용 2013. 5. 15].
10) D. Soergel, *op. cit.*, p.126.
11) Government of Canada Core Subject Thesaurus, 〈http://www.thesaurus.gc.ca/default.asp?lang=En&n=0073D232-1〉 [cited 2013. 5. 15].
12) J. Breuker, A. Valente, and R. Winkels, "Use and Reuse of Legal Ontologies in Knowledge Engineering and Information Management." In: *Law and the Semantic Web*, edited by Benjamins et al. (Berlin Heidelberg : Springer-Verlag, 2005), pp.36-64.
13) D. Soergel, op. cit., p.126.

캐나다 코어 주제 시소러스는 캐나다 정부의 정보 자원을 다루는 모든 분야를 표현하는 용어로 구성된 영어와 프랑스어의 이중 언어 시소러스이다. 이 시소러스는 모든 영역의 지식에 대하여 다양한 계층을 위해 만들어졌으며, 전문용어보다는 일반적인 용어로 구성되어 있다.[14]

한편, 온톨로지는 톱 레벨 온톨로지, 코어 온톨로지, 도메인 온톨로지로 나눌 수 있다. Breuker는 법률 코어 온톨로지인 LRI-Core 온톨로지를 구축 및 활용함에 있어서, 상위-코어-도메인의 차례로 계승하는 모델을 제시하였으며 코어 온톨로지는 상위 온톨로지와 도메인 온톨로지의 앵커 기능을 하도록 하였다. LRI-Core 온톨로지는 상위 개념과 코어 개념을 포괄하고 있으며 코어 온톨로지를 앵커로 하여 네덜란드의 형법 온톨로지인 OCL.NL을 작성하였다.[15]

본 연구에서 말하는 코어 시소러스는 톱 레벨의 용어를 포함한 전 주제를 포괄할 수 있는 시소러스로 사용하였다. 또한, 매크로시소러스와 마이크로시소러스를 상정했을 때 특정 주제 분야의 마이크로시소러스를 부분으로 가지는 매크로시소러스의 최상위 시소러스를 코어 시소러스라고 하였다.

2. 매크로시소러스와 마이크로시소러스

Lancaster에 의하면 마이크로시소러스라고 하는 용어는 일반적으로 받아들여지는 정의는 없고, 종종 보다 작은 그리고 분화된 시소러스에 관계되는 것에 토대로 하여 사용되어진다고 정의하였다. 마이크로시소러스는 보다 큰 시소러스로부터 추출된 용어들의 특별한 목적을 위해 분화된, 그러므로 호환성이 있는 부분집합을 말한다. 그러므로 마이크로시소러스는 어떤 더 넓은 시소러스에 매핑되는 분화된 어휘로서 정의될 수 있다. 그리고 그 시소러스의 계층구조에 완전히 포함되어진다고 하였다.[16]

또한 그는 완전히 새로운 시소러스를 구축하기 전에 각자의 요구에 맞는 현존하는 시소러스를 검토할 필요성을 제기하면서, 경우에 따라서는 다양한 전문 영역으로 시소러스를 전개하는 하나의 형식으로 보다 범용성이 있는 시소러스의 구조에 맞춘 마이크로시소러스의 구축을 제안하였다. 예를 들면, 기존의 의학 시소러스의 계층구조에 맞춘 당뇨병 시소러스의 신설 등이다.[17]

예를 들어, 건축 재료 중 타일을 주로 다루는 시소러스에서 건축 관련 전반적인 주제어를 모두 채택하는 것은 합리적이지 못하다. 기존에 건축 재료 시소러스가 존재한다면, 타일 관련 용어는 기존의 것에 그 부분만 확장하여 사용하는 것이 효율적일 것이다.

Aitchison et al.은 통제된 언어들 사이에 구축된 호환성이 있는 마이크로시소러스, 즉 세분화된 시소러스가 보다 넓은 시소러스인 매크로시소러스의 계층적 구조위에 매핑되거나 전체적으로 그

14) Government of Canada Core Subject Thesaurus, *op. cit.*
15) J. Breuker, A. Valente and R. Winkels, *op. cit.*, p.48.
16) F. Wilfrid Lancaster, *op. cit.*, pp.198-202.
17) F. Wilfrid Lancaster, 情報システムのための構築と利用. 松村多美子. 鈴木祐滋 譯(東京 : 情報科學技術協會. 1989), p.6.

안에 통합될 때 매크로시소러스에 의지할 수 있다고 하였다.[18]

Willpower Information는 마이크로시소러스의 항목에서 "마이크로시소러스는 대개 전체 시소러스의 부분보다 작은 주제 영역의 용어를 포함하는 시소러스의 부분집합이다"라고 기술하고, 그 예시로서 유네스코 시소러스는 7개의 마이크로시소러스로 세분화된다고 하였다. 유네스코 시소러스를 기반으로 하는 영국기록보관소 시소러스(UK Archival Thesaurus)는 지식의 영역이라 불리는 마이크로시소러스로 세분화되고 있다고 기술하고 있다.[19]

한편, 국내에서의 매크로시소러스의 정의는 망라적으로 많은 양의 시소러스로 정의하는 경향이 있다.

장명희는 한국어 매크로시소러스는 국어의 모든 어휘를 대상으로 하고 어휘들 사이의 관계를 규명한 어휘 의미 분류집이라고 하였다. 대상이 되는 분야의 범위에 따라 매크로시소러스와 마이크로시소러스로 나뉘며, 광범위하고 복합적인 분야를 망라하는 시소러스를 매크로시소러스라 하고, 좁은 분야나 특정의 분야의 시소러스를 마이크로시소러스라 한다고 하였다. 해당 논문에서는 시소러스는 국어만을 대상으로 하며 국어에 있는 모든 어휘를 작업 대상으로 하여 "한국어 매크로시소러스"라는 명칭을 사용하였다. 한국어 매크로시소러스는 광범위하고 복합적인 분야를 망라하는 매크로시소러스인 것이다.[20]

최석두는 한글 매크로시소러스를 개발할 때는 전통적인 시소러스에 더하여 다음과 같은 부분이 보완되어야 한다고 생각하였다. 첫째, 용어를 망라해야 한다. 둘째, 관계를 확장하여야 한다. 셋째, 각종 언어정보를 가져야 한다. 넷째, 해설을 가져야 한다고 지적하였다. 예상으로는 등록용어수가 20만어 정도가 되면 일단 매크로시소러스의 형식을 갖추게 될 것이며, 포괄적인 분야에서의 사용도 가능하리라 생각하였으며, 궁극적으로는 100만 단위 규모의 용어를 가질 때 명실 공히 매크로시소러스가 되리라 생각된다고 하였다.[21]

본 연구에서는 Lancaster, Aitchison et al.의 정의를 따른다. 매크로시소러스와 마이크로시소러스는 상대적 개념이다. 많은 용어를 포함하고 있는 것만으로 매크로시소러스가 아니라 마이크로시소러스를 부분집합으로 가지고, 매크로시소러스와 마이크로시소러스호환성이 있을 때에 성립하는 것으로 하였다.

3. 선행 사례 분석

현존하는 뉴스 영역의 시소러스는 주로 일간종합신문이나 경제종합신문 등의 뉴스정보를 관리하기 위해 만들어진 경우가 많다.

---

18) J. Aitchison, A. Gilchrist and D. Bawden, *op. cit.*, p.177.
19) Willpower Information, 〈http://www.willpowerinfo.co.uk/glossary.htm〉 [cited 2013. 5. 15].
20) 장명희, 한국어 매크로 시소러스의 작성법 연구(석사학위논문, 숙명여자대학교 대학원 국어국문학과, 2001), p.12, p.23.
21) 최석두, "한글 매크로시소러스 구축의 실제," 한국정보관리학회 학술대회 논문집, 제5호(1998), pp.223-226.

시소러스는 종합지를 대상으로 하는 국내의 뉴스ML 시소러스가 있고, 경제신문을 대상으로 하는 경제신문시소러스가 있다. 미국의 경우 뉴욕타임스 시소러스와 워싱턴 포스트 시소러스가 유명하다. 또한, 일본의 경우는 종합지의 "뉴스시소러스"와 경제종합의 "일본경제시소러스", 공업신문의 "일간공업시소러스" 등이 있다. 뉴욕타임스 시소러스 외에는 모두 키워드를 카테고리별로 나누고 있다.

본 연구에서 용어 수집과 용어 구조화에 참고한 한국언론재단(현 한국언론진흥재단)의 뉴스ML 시소러스를 상세히 분석하면 다음과 같다.

뉴스 ML 시소러스는 다양한 매체, 다양한 형태의 뉴스에 널리 적용할 수 있도록 뉴스의 주제가 되는 중요한 어휘들을 수집하여 동등관계, 상하관계, 연관관계 등을 체계화한 뉴스 종합시소러스로서, 뉴스 정보관리 시스템에서 사용할 목적으로 작성된 것이다. 어느 특정 주제에 한정하지 않고 뉴스를 구성하고 있는 전 분야를 대상으로 하였으며, 뉴스를 구성하고 있는 넓은 영역 중에서 뉴스정보를 관리하고 검색하기 위해 필요한 개념 중 일반 주제명, 중요 기관단체명과 지명을 수록하고 인명, 회사명 등은 수록하지 않았다. 용어의 총 수는 26,782어이고 이 중 우선어는 21,345어, 비우선어는 5,437어이다.[22)]

국내외의 주요 뉴스 관련 시소러스의 개황을 〈표 1〉에 나타내었다.

〈표 1〉 국내외 주요 뉴스 영역의 시소러스 개황

시소러스	작성자	대상 영역	구성
뉴스ML 시소러스(2008)	한국언론재단	뉴스 종합 신문	26,782어(21,345어+5,437어)
ニュース·シソーラス(2004)	廣木守雄服部信司	뉴스 종합 신문	14,289어(10,380어+3,909어)
경제신문시소러스(1993)	한국경제신문사	경제 종합 신문	13,646어(10,406어+3,240어)
日経シソーラス(2013)	日経텔레콤21	경제 종합 신문	약13,000어*
Washington post thesaurus([1989])	The Washington post Company	뉴스 종합 신문	1,856어(비우선어는 없음)
New York Times Descriptors	The New York Times Company	뉴스 종합 신문	498어만 링크드 데이터로 공개
日刊工業シソーラス 제3판(1989)	日刊工業新聞社	일간 공업 신문	13,003어(12,450어+553어)

* 1989년 책자형에는 15,434어라고 명시하였으나 현재는 온라인상에서만 관리되며 수시로 추가 및 삭제하여 2006년에는 약13,000어라고 하고 있다[23)]

기타 영역의 본 연구에 영향을 준 선행 시소러스들의 사례를 이하에 설명한다.

Tesqual은 ISO 2788-1985와 Aitchison et al.의 구축 절차에 따라 설계되었으며 고등 교육의 품질 관리의 영역에서 점점 커져가는 문제를 해결하기 위하여 제어된 언어의 필요성에 따라 개발되었다. 이것은 대학교육과 관련된 이용자들의 기대와 필요를 충족시키기 위한 것이며 마이크로시소러스의 형식을 취하고 있다. 마이크로시소러스 Tesqual은 동의, 상하, 관련 관계를 가지고 있으며 이는 세부적인 문헌들의 내용을 정의하고 개념화하는 "핵심" 어휘를 사용하는 일반적인 이용자들

22) 한국언론재단, 뉴스ML 시소러스(상)(하)(서울 : 한국언론재단, 2008), p.i.

23) Nikkei Telecom, 〈http://www.nikkeitel.com/〉 [cited 2013. 5. 1].

과 학생, 교육 전문가, 조사자, 과학자들을 겨냥한다. 최종 목표는 전문가들이 특정 정보시스템에서 오는 이러한 문헌들을 저장하고 찾아내는 것을 돕는 것이다. Tesqual은 9개의 일반적인 의미론적 영역으로 나누어지는데 이들 영역은 2,425개의 용어들로 구성되어 있고, 우선어 2,013개, 비우선어 412개이다. 9개의 의미론적 영역은 또한 더욱 중간 정도의 60영역과 그 하위 영역인 세부적인 영역으로 세분화된 마이크로시소러스로 구성되어 있다[24].

EUROVOC은 유럽공동체 조직의 문헌정보를 처리하기 위해 특별히 고안된 시소러스이며, 다국적 성격의 공동체적 관점과 개별 국가적 관점에 대해서도 다루고 있다.[25] EUROVOC의 구성은 영어를 기준으로 21개의 주제 분야, 127개의 마이크로시소러스, 6,883개의 우선어, 8,343개의 비우선어, 802개의 스코프 노트, 233개의 정의, 58개의 히스토리 노트로 구성되어 있다.[26]

OECD의 MACROTHESAURUS는 경제와 사회개발 영역의 정보처리를 위한 시소러스로서 1969년 초판 이래 현재 1998년 제5판이 공표되었다. 제5판은 상위 분류 19개, 하위 분류 124개의 마이크로시소러스로 구성된 5,174어의 우선어와 2,055어(영어 873어, 프랑스어 571어, 스페인어 613어)의 비우선어로 구성된 시소러스이다.[27]

한편, 캐나다 코어 시소러스는 모두 19개의 카테고리를 가지고 있다. 캐나다 코어 시소러스의 키워드는 모두 최대 4개까지의 카테고리가 분류되어 있으며, 영어 우선어 2,195어, 프랑스 우선어 2,198어가 등록되어 있다.[28]

이들의 선행 사례를 정리하면 〈표 2〉와 같다.

〈표 2〉 뉴스 영역 이외의 선행 사례의 분석

시소러스	공표자	마이크로시소러스	우선어	영역	언어	비고
Tesqual	Espacio Europeo de Educación Superior	상위: 9 하위: 60	2,013어	대학의 품질 관리	스페인어	모든 우선어를 계층으로 연결
EuroVoc	EU	상위: 21 하위: 127	6,883어	EU 주요 활동	23 EU 언어	
MACROTHESAURUS	OECD	상위: 19 하위: 124	5,174어	경제와 사회 개발	영어, 프랑스어, 스페인어	
Canada Core	캐나다 정부	(19개의 카테고리)	2,195어	정부문헌	영어, 프랑스어	프랑스어의 우선어는 2,198어

24) M. M. Aranda, Tesqual: *A Microthesaurus for Use in Quality Management in European Higher Education*, 2010. [cited 2013. 6. 20].

25) 한국과학기술정보연구원, 한·영·일 대역 과학기술분류표, 시소러스, 용어사전 개발방안 연구(서울 : 한국과학기술정보연구원, 2001) pp.30-32.

26) EUROVOC Home page, 〈http://eurovoc.europa.eu/drupal/〉 [cited 2013. 5. 1].

27) OECD Development Centre, *Macrothesaurus for Information Processing in the Field of Economic and Social Development*, 5th Ed. 1998. [cited 2013. 6. 20].

28) Government of Canada Core Subject Thesaurus, *op. cit.*

Ⅲ. 뉴스 코어 시소러스의 구축

1. 구축 방안

ISO 2788[29]과 Aitchison et al.[30]의 기준과 절차에 따라 구축한다. 이를 바탕으로 뉴스 코어 시소러스 구축을 위한 설계 원칙을 제시한다. 뉴스 코어 시소러스는 마이크로시소러스와 링크를 고려하여 매크로시소러스로서의 역할을 하고 마이크로시소러스는 그것의 완전한 부분집합이 될 수 있게 한다. 즉, 독립하여 사용할 수 있고, 마이크로시소러스와 연계하여 병합 및 통합된 시소러스로 사용할 수 있도록 한다. 뉴스 관련 영역의 전 범위를 커버하고 망라성을 확보하기 위해 종합지 위주의 분류표 및 시소러스를 참조하였다.

가. 설계 원칙

본 연구에서의 뉴스 코어 시소러스의 구축은 다음과 같은 설계 원칙을 설정하였다.

첫째, ISO 2788을 준수하고, Aitchison et al.의 구축 절차를 따른다.

둘째, 뉴스ML 시소러스를 위주로 용어를 수집하고 구조를 참조하되, 문화체육관광부와 각 시도에 등록된 신문 목록을 참조하여 위로는 코어 시소러스를 아래로는 마이크로시소러스를 만든다.

셋째, 코어 시소러스는 망라성을 고려해야 하기 때문에 가능한 한 넓은 영역을 커버할 수 있는 용어를 대상으로 한다.

넷째, 더미텀(dummy term)을 사용하여 가능한 한 관련 용어를 모은다. 더미텀이란, 색인시 문헌에 할당되지 않는 형식적인 용어이다.[31] Aitchison et al.[32]은 스코프 노트는 더미텀을 나타내기 위해서 사용할 수 있다고 하였다. 더미텀은 체계 표시의 구조를 명확하게 하기 위해 필요한 용어이지만, 그것 자체는 색인 작성에 적당하지 않은 것들이다. 이 용어는 자모순 표시에서 분류기호(classmark, 예를 들면, 아래 예시의 BV)를 붙인다. 본 연구에서 분류기호는 모든 용어에 부여하므로 각각 처리하되, 더미텀은 공통의 「00.01」을 일괄 부여하여 관리한다. Aitchison et al.이 제시한 사례를 다음에 나타낸다.[33]

예: EDUCATION OF SPECIFIC CATEGORIES OF STUDENTS BV

29) ISO 2788, 시소러스 개발 지침, 최석두, 정동열 공역(서울 : 문헌정보처리연구회, 1994).
30) J. Aitchison, A. Gilchrist and D. Bawden, *op. cit.*, pp.145-167.
31) ISO 2788, *op. cit.*, p.13.
32) J. Aitchison, A. Gilchrist and D. Bawden, *op. cit.*, p.35.
33) *Ibid.*

SN Do not use as an indexing term.
NT Exceptional student education
Parent education
Women's education

다섯째, 동등관계 중 업워드 포스팅(upward posting)은 사용하지 않으며, 유사 동의어는 최대한 배제한다. 업워드 포스팅은 본래 하위 개념이지만, 넓은 주제를 대상으로 하는 시소러스에서 상위어에 부가하여 비우선어로 인정되는 것이다. 뉴스 코어 시소러스는 매크로시소러스와 마이크로시소러스를 전제로 하고 있기 때문에 마이크로시소러스에서는 우선어가 될 수 있는 업워드 포스팅은 하지 않는다. 유사 동의어는 일반적으로 의미가 다르지만 색인 작성에 있어서 동등관계로 처리하는 용어이며, 의미가 현저하게 중복되는 용어를 포함한다. 그러나 주변 영역에서만 사용해야 한다고 국제규격에서는 설명하고 있다.[34] 뉴스 코어 시소러스에서는 주변 영역은 다루지 않고 있으며, 마이크로시소러스와 연계하여 사용하는 경우를 상정하였기 때문에 마이크로시소러스에서는 유사 동의어를 채택하더라도 코어 시소러스에서는 최대한 배제하는 것으로 한다.

여섯째, 카테고리를 설정하여 매크로시소러스와 마이크로시소러스가 될 수 있도록 제시한다.

일곱째, 수록 용어의 종류는 일반주제명, 기관단체명, 지역명 등 모든 종류를 포함하는 것으로 한다. 단, 인명, 상품명, 회사명은 제외한다.

여덟째, 시소러스의 구성은 자모순, 계층순, 카테고리별 키워드 일람 등으로 구성하며, 시소러스 내에서 사용하는 기호는 국제규격 ISO 2788에 준하여, SN(Scope Note), UF(Used For), TT(Top Term), BT(Broader Term), NT(Narrower Term), RT(Related Term)를 사용하고, 카테고리는 CA(Category)를 사용한다.

아홉 번째, 색인어의 띄어쓰기는 실시하지 않으며, 동형이의어의 식별은 키워드의 말미에 괄호를 부가하고 식별구를 넣는다. 식별구는 상위어나 카테고리명을 넣는다. 용어의 일부가 되는 동형이의어의 식별자인 괄호와 용어의 중간에 존재하는 중간점은 마침표(.)를 사용한다.

열 번째, 자모순 배열은 숫자, 한글, 영문 두문자 순이며, 한글은 가각까 순으로 하고, 지시기호의 배열순서는 SN, CA, UF, TT, BT, NT, RT 순으로 한다.

나. 매크로시소러스와 마이크로시소러스의 구성 체계

문화체육관광부의 '정기간행물등록현황(2012)'[35]과 포털 사이트 다음의 'Daum 디렉토리',[36] 그

34) ISO 2788, *op. cit.*, p.50.
35) 문화체육관광부(정기간행물등록현황) 〈http://www.mcst.go.kr/web/s_data/deptData/deptDataView.jsp?pSeq=81〉 [인용 2013. 5. 3].
36) Daum 디렉토리 〈http://directory.search.daum.net/site_list.daum〉 [인용 2013. 5. 15].

리고 각종 뉴스 관련 분류표와 시소러스를 참조하여 카테고리를 설정하였다.

'정기간행물등록현황'은 문화체육관광부와 각 시도에 등록된 정기간행물의 현황으로서 본 연구에서는 일간신문, 주간신문, 특수신문, 인터넷신문 6,800여 종의 목록 내의 '성격'을 참조하였으며, 'Daum 디렉토리'의 경제신문, 스포츠·연예신문, 종교신문 등 35개 항목을 참조하였다.

기사표준분류표는 대분류 10항목과 중분류 78항목, 소분류 594항목이며, IPTC 뉴스 코드는 17개의 대분류와 3계층의 총 1,397항목이며,[37] 아사히 기사데이터베이스 분류표는 대분류 10항목, 중분류 92항목, 소분류 669항목으로 구성되어 있다.[38]

한편, 뉴스ML 시소러스는 8개의 카테고리를 41개로 세분화하고 있으며, 일본의 뉴스시소러스는 7개의 카테고리를 42개로 세분화하고 있다. 또한 워싱턴 포스트 시소러스는 모든 키워드를 알파벳 기호를 사용하여 1,856어를 계층으로 구성하고 있다.

참조한 뉴스 주제분류표와 시소러스의 카테고리표의 상위 구조를 〈표 3〉에 나타내었다.

〈표 3〉 참조한 카테고리

분류표						시소러스					
기사표준분류		IPTC 뉴스코드		아사히분류		뉴스ML		워싱턴포스트		ニュースシソーラス	
0	총류	01	예술, 문화와 예능	0	총류	A	정치	A	정부와 정치	A	정치
1	정치	02	범죄, 사법	1	정치	B	경제	B	군사문제	B	경제
2	경제	03	재해와 사고	2	경제	C	산업	C	비즈니스, 상업, 경제적 상황	C	사회
3	산업	04	경제, 비즈니스와 제정	3	노동	D	사회	D	직업, 노동, 고용	D	사건
4	사회	05	교육	4	문화	F	범죄	E	운수	E	문화
5	사건·사고	06	환경문제	5	과학	G	문화	F	커뮤니케이션과 정보	F	국제
6	문화	07	건강	6	사회	H	자연	G	도시개발, 빌딩과 구조	G	공통
7	과학	08	인간적 흥미	7	사건	J	국제	H	에너지와 환경		
8	스포츠	09	노동	8	스포츠			J	사고와 재해		
9	국제	10	생활과 레저	9	국제			K	사회조직과 관심사(범죄)		
		11	정치					L	법과 입법		
		12	종교와 사상					M	교육		
		13	과학과 기술					N	철학, 종교, 초자연		
		14	사회문제					P	예술과 예능		
		15	스포츠					G	레저		
		16	사회불안, 갈등과 전쟁					R	경쟁적 스포츠		
		17	기상					S	사람과 인구		
								T	인체, 행태, 건강		
								U	가족경제		
								V	소비재		
								W	과학과 기술		
								X	생물		
								Y	시간/역사		
								Z	지리적 용어		

37) IPTC Home page, 〈http://cv.iptc.org/newscodes/subjectcode/〉 [cited 2013. 5. 15].

38) 朝日新聞社ニューメディア本部, 朝日記事データベース分類の手引き(東京 : 朝日新聞社ニューメディア本部, 1989).

상하위 카테고리는 매크로시소러스와 마이크로시소러스의 관계에 있으며, 일간종합신문을 대상으로 하는 신문 종합, 일간경제신문을 대상으로 하는 경제 종합, 스포츠신문을 대상으로 하는 스포츠/연예, 그리고 사회, 문화, 자연, 국제 영역의 각각을 대상으로 하는 사회 종합, 문화 종합, 자연 종합, 국제 종합 등으로 마이크로시소러스를 구성할 수도 있다.

카테고리의 기호법은 아라비아 숫자 두 자리씩(00.00)으로 구분한다. 점은 가독성을 위한 의미 없는 기호이다. 이렇게 완성된 카테고리는 대분류 15항목, 중분류 87항목이며, 〈표 4〉에 나타내었다.

〈표 4〉 완성된 카테고리

기호	카테고리명	기호	카테고리명	기호	카테고리명	기호	카테고리명
00	공통	02.09	IT・통신	06.01	환경공해	11	스포츠
00.00	공통일반	02.10	전기・전자	07	생활	11.00	스포츠일반
00.01	더미텀	02.11	상업	07.00	생활일반	11.01	종합경기
00.02	국내지역	02.12	무역	07.01	의	11.02	육상
00.03	국외지역	03	사회	07.02	식	11.03	구기
01	정치	03.00	사회일반	07.03	주	11.04	수상
01.00	정치일반	03.01	복지	08	사건사고	11.05	동계
01.01	대통령	03.02	가정	08.00	사건사고일반	11.06	격투기
01.02	의회	03.03	아동	08.01	범죄	11.07	레포츠
01.03	선거	03.04	청소년	08.02	사고	12	학문
01.04	행정	03.05	여성	08.03	자연재해	12.00	학문일반
01.05	재정	03.06	노인	09	문화	12.01	인문학
01.06	사법	03.07	레저	09.00	문화일반	12.02	사회과학
01.07	법률	03.08	행사	09.01	문화재	12.03	자연과학
01.08	외교	04	교육	09.02	문학	12.04	공학
01.09	국방	04.00	교육일반	09.03	출판.인쇄	13	자연
02	경제	04.01	육아.유치원	09.04	종교	13.00	자연일반
02.00	경제일반	04.02	초중고교육	09.05	역사	13.01	기상
02.01	금융	04.03	대학교육	10	예술	13.02	생물
02.02	기업・경영	04.04	성인교육	10.00	예술일반	13.03	우주
02.03	노동	05	건강	10.01	음악	14	국제
02.04	농림수산	05.00	건강일반	10.02	미술	14.00	국제일반
02.05	공업	05.01	의료	10.03	무용	14.01	국제기구
02.06	자원・에너지	05.02	질병	10.04	연극		
02.07	건설	06	환경	10.05	영화		
02.08	운수・교통	06.00	환경일반	10.06	연예		

2. 용어의 수집

종합적인 뉴스 영역 분류표 및 시소러스에서 수집하고 분석・정리하였다.

가. 용어의 수집 방법

용어 수집원은 뉴스의 종합을 다루는 분류표로서 IPTC News Codes, 기사자료표준분류표, 아사히 기사데이터베이스 분류표(이하 아사히 기사분류표라 한다)의 모든 항목과 시소러스로서 뉴스ML 시소러스, 일본의 뉴스시소러스 등의 최상위 용어 및 고립어를 망라하여 조사하였으며, 특히 뉴스ML 시소러스의 순열색인(Permuted Index, KWOC 색인)을 정리한 후 어휘 조각(fragment)도 수집하였다. 더미텀의 선정 기준은 주제어로서는 의미를 이루지 못하지만 관련 키워드를 한 곳으로 모을 수 있는 용어라고 판단된 경우이며, 주로 위의 뉴스ML 시소러스의 순열색인에서 취했다.

나. 용어 수집원

용어의 수집원은 영역을 한정하지 않는 종합지를 대상으로 하는 시소러스나 분류표를 대상으로 하였다.

분류표로서는 IPTC(International Press Telecommunications Council, 국제언론통신협의회)가 공표한 News Subject Codes, 국내의 '기사자료 표준분류표', 일본의 아사히 기사분류표를 활용하였다.

기존의 뉴스 관련 시소러스에 대해서는 "II. 3. 뉴스 영역의 시소러스"에서 검토한 바와 같다.

용어 수집원인 분류표 및 시소러스는 〈표 5〉와 같고, 총 어휘 수와 수집 대상이 되는 어휘 수를 기술하였다.

〈표 5〉 용어 수집원의 개요

분류표 또는 시소러스	총 어휘 수	수집 대상 어휘 수	비고
IPTC News Codes(2013)	17개의 대분류와 3계층의 총 1,397항목	1,397항목	전 분류 항목
뉴스ML 시소러스(2008)	26,782어(21,345어+5,437어)	2,603어	최상위 및 고립어 모두
ニュース·シソーラス(2004)	14,289어(10,380어+3,909어)	4,608어	최상위 및 고립어 모두
Washington post thesaurus([1989])	1,856어	1,856어	비우선어는 없음.
New York Times Descriptors(2013)	498어만 링크드 데이터로 공개	498어	인물, 조직, 지명 등은 별도 공개
기사표준분류표 제2판(1991)	총 대분류 10항목, 중분류 78항목, 소분류 594항목	594항목	세분류는 생략
아사히 기사분류표(1989)	대분류 10항목, 중분류 92항목, 소분류 669항목	1,300항목	항목에 포함된 색인어 및 별도 관리되고 있는 지명, 기사 유형은 제외

3. 용어의 통제

더미텀은 활용 면에서 색인어로서의 사용을 금지함으로써 마이크로시소러스의 영역에서의 혼란을 방지하고, 새로운 색인어를 생성하여 사용할 수 있도록 안내하는 역할을 한다. 예를 들면, 축산신문에서 “시설”은 당연히 “축산시설”을 지칭할 것이다. 이 때 색인어는, “축산시설”로 하고 “시설”로 하는 것을 금지하는 것이다. 더미텀은 스코프노트에 “색인어로서는 사용하지 않는다”라고 지시하였다. 더미텀은 유사한 개념을 한 곳에 모으는 역할도 하지만 마이크로시소러스와 병합 또는 통합을 위해서도 고려하였다.

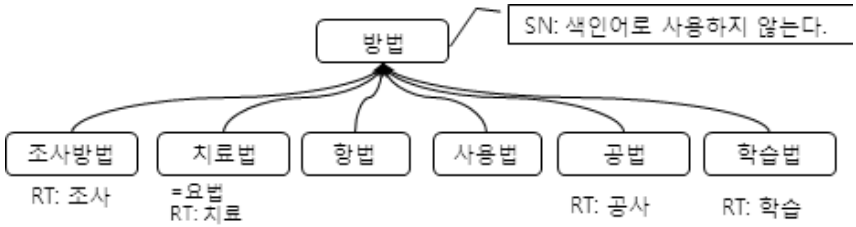
4. 용어의 구조화

주로 더미텀과 최상위어를 사용하여 카테고리별로 클러스터링한 후 용어를 구조화하였다. 예를 들어, 최상위어로 되어 있는 ‘조사방법’을 사용하여 더미텀 ‘방법’을 선택하고 모든 최상위어를 조사하여 ‘방법’에 해당하는 것을 모았다.

용어의 구조화는 미들 아웃방식을 취하거나 최상위에 해당하는 더미텀을 이용하여 톱다운으로 설정하였으며 마이크로시소러스 카테고리를 고려하여 깊이를 조절하였다.

동등관계에 있어서 업워드 포스팅은 채택하지 않았으며, 유사 동의어는 최대한 배제하였다. 상하관계는 마이크로시소러스에서 확장을 고려하기 때문에 널리 쓰이는 중요한 키워드가 아니면 앵커로서 기능할 수 있는 단계까지만 설정하였다.

뉴스 코어 시소러스의 모든 정보를 용어 구조도를 사용하여 나타냈으며 사례를 〈그림 1〉에 나타내었다.



〈그림 1〉 용어의 구조도 예시

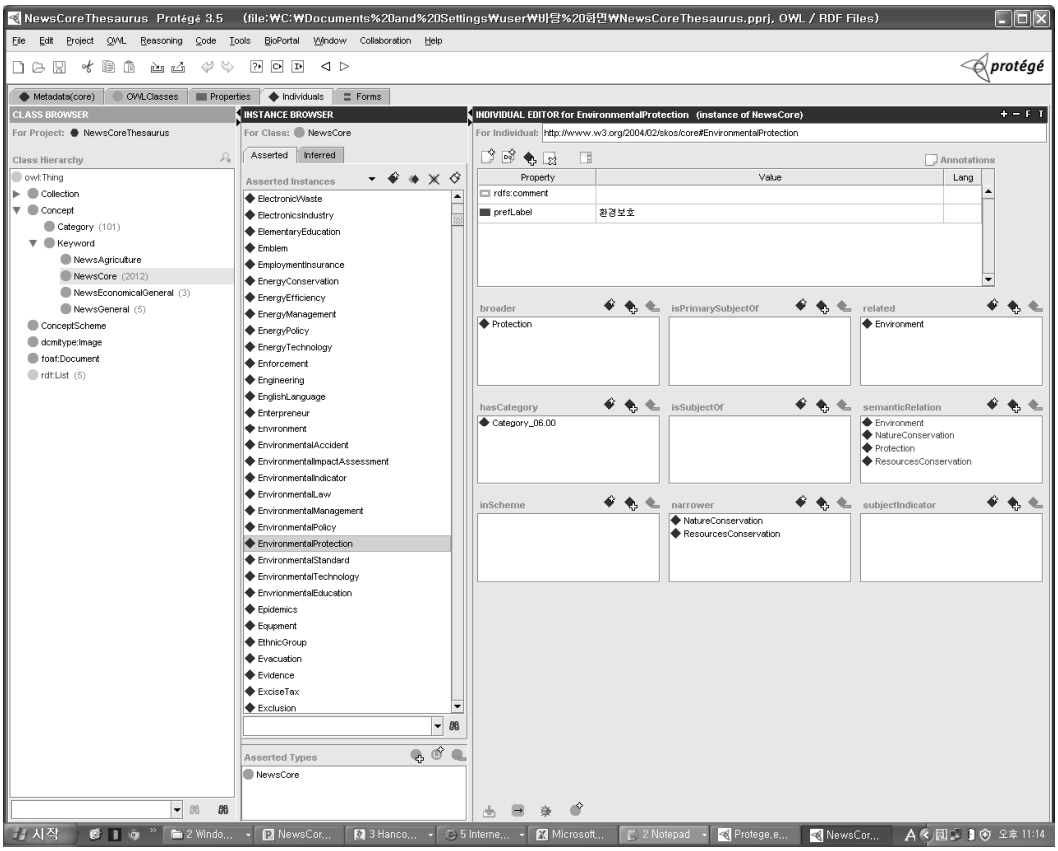
5. 완성 및 통계

가. SKOS의 완성

본 연구에서는 기존의 skos-core-owl-dl.owl[39]을 Protégé 3.5에 임포트하여 작성하였다. “Concept”의 하위에 “Keyword”와 “Category”를 신설하고 “Keyword”의 하위에 다시 “NewsCore”, “NewsGeneral”, “NewsEconomicalGeneral”과 같은 형식으로 클래스를 생성하고 각각의 클래스에 인스턴스를 생성하였다.

hasKeyword↔hasCategory의 프로퍼티를 사용하여 카테고리와 키워드와의 링크를 하고, broader↔narrower, related를 사용하여 키워드 간의 관계를 지시하였으며, 스코프 노트는 scopeNote를 사용하여 기술하였다. 카테고리 간에도 broader↔narrower을 형성하였다. 단, 카테고리 간의 ‘related’는 생성하지 않았다.

그 입력 사례를 〈그림 2〉에 나타내었다.



〈그림 2〉 SKOS의 입력 예시

39) W3C, 〈http://w3.org/2004/02/skos/core/owl-dl/skos-core-owl-dl.owl〉 [cited 2013. 5. 15].

나. 통계

결과로서 얻어진 뉴스 코어 시소러스의 통계는 〈표 6〉과 같다.

〈표 6〉 뉴스 코어 시소러스의 통계

용어	합계	참고
우선어	2,012어	
비우선어	74어	
마이크로시소러스	101분야	대분야 15, 중분야 86
더미텀	117어	

구축된 뉴스 코어 시소러스는 뉴스의 넓은 주제 영역을 가지는 용어를 코어 시소러스에서 커버하고 마이크로시소러스에서 확장하고 있기 때문에 넓은 영역을 커버할 수 있을 것이다.

Ⅳ. 뉴스 코어 시소러스의 활용 방안

뉴스 코어 시소러스의 구축의 주 목적은 신문의 모든 영역의 뉴스 정보를 관리하기 위한 매크로 시소러스의 구축이므로 단독으로 사용할 수도 있지만 부분집합으로 마이크로시소러스의 사용을 전제로 하고 있다. 본 연구에서 뉴스 코어 시소러스의 활용을 네 가지로 나누어 검토하였다. 첫째, 뉴스 코어 시소러스의 용어 구조를 더욱 확장하여 활용하는 방안, 둘째, 각각의 마이크로시소러스를 사용하는 영역의 뉴스 정보를 병합하거나 통합하여 사용하는 방안, 셋째, 이상의 조건을 전제로 하지 않고 단독으로 사용하는 방안, 넷째, 마이크로시소러스의 카테고리를 활용하는 방안이다. 마지막으로, 실제 이용자의 입장에서 공유와 상호운용성을 위한 활용 방안을 제시하였다.

1. 용어 구조의 확장

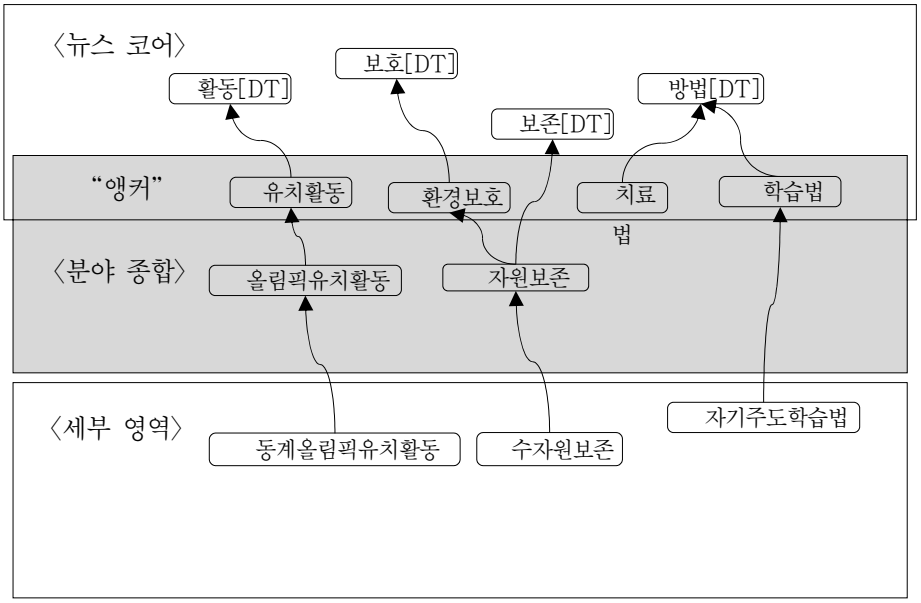
뉴스 코어 시소러스의 용어 구조를 그대로 확장하는 방안과 마이크로시소러스를 구축하여 확장하는 방안으로 나눌 수 있다.

가. 용어 구조의 확장

뉴스 코어 시소러스의 모든 구조를 그대로 사용하되 관련 용어를 더욱 확장해가는 방안이다. 소규모 지역 신문의 경우 뉴스 주제 영역은 뉴스 코어 시소러스와 같이 넓은 주제를 차지하지만, 정보

량이 적은 경우를 상정한 활용 방안이다.

예를 들면, 뉴스 코어 시소러스의 '방법'과 치료법'이 있고, 여기에 '민간요법'을 신설하는 방법이다. 〈그림 3〉에 확장에 대한 개념을 나타내었다.



〈그림 3〉 용어의 구조를 확장하는 개념도[40]

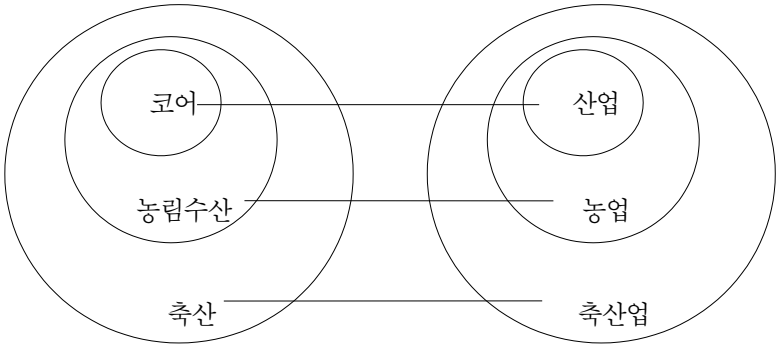
나. 마이크로시소러스로의 확장

주제 영역이 특수하고 정보량이 적은 소규모의 신문을 상정한 활용 방안이다. 신규로 마이크로시소러스를 만들어 뉴스 코어 시소러스는 그대로 두고 이를 앵커로 하여 확장하는 방안이다. 각각의 소규모 신문이나 특수 분야를 다루는 신문들은 뉴스 코어 시소러스를 공통으로 자신의 신문을 마이크로시소러스로 확장할 수 있다.

예를 들어, 우리나라의 농업 관련 신문은 '농민신문', '농수축산신문', '농업축산신문', '축산경제신문', '한국농어민신문', '여주농민신문', '농업인신문', '축산신문' 등이 있다. 이들은 경제종합에 대한 마이크로시소러스를 형성하고 있다. 이들의 영역을 '농림수산'으로 하고, 뉴스 코어 시소러스, 경제종합 마이크로시소러스, 농림수산 마이크로시소러스의 순으로 확장해갈 수 있다. 가령, '축산신문'을 대상으로 뉴스 기사를 색인 작성을 하는 경우, '보호←환경보호←자원보호←종돈보호'와 같이 '종돈보호'를 기존의 '자원보호'의 하위개념으로 등록할 수 있다. 앵커가 될 수 있는 용어가 존재하지 않는다고 판단된 경우는 카테고리를 확인하고 신설할 수 있다.

40) J. Breuker, A. Valente and R. Winkels, *op. cit.*, p.48을 참조함.

이 유형의 사례를 〈그림 4〉에 나타내었다.



〈그림 4〉 마이크로시소러스를 사용하여 확장하는 개념도

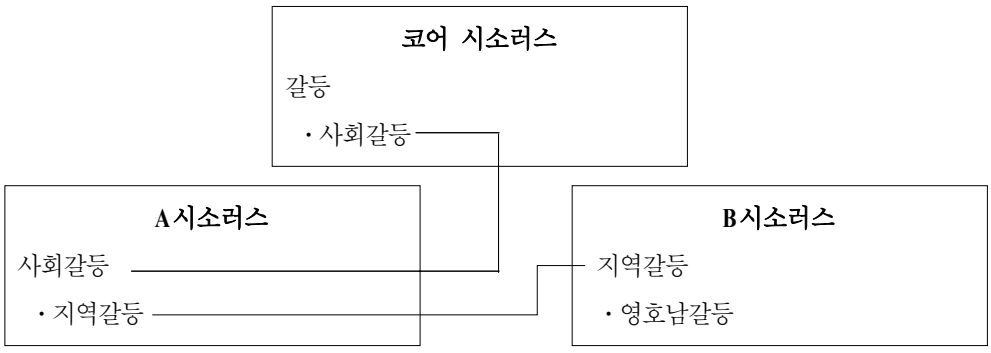
2. 병합 또는 통합

기존의 시소러스를 뉴스 코어 시소러스를 매크로시소러스로 하여 병합 또는 통합하는 활용 방안이다.

가. 병합

기존의 뉴스 관련 시소러스를 사용하고 있는 경우 뉴스 코어 시소러스와 연계하되, 둘 다 존속하면서 활용하는 방안으로, 뉴스 아카이브와 같은 곳에서 활용할 수 있다.

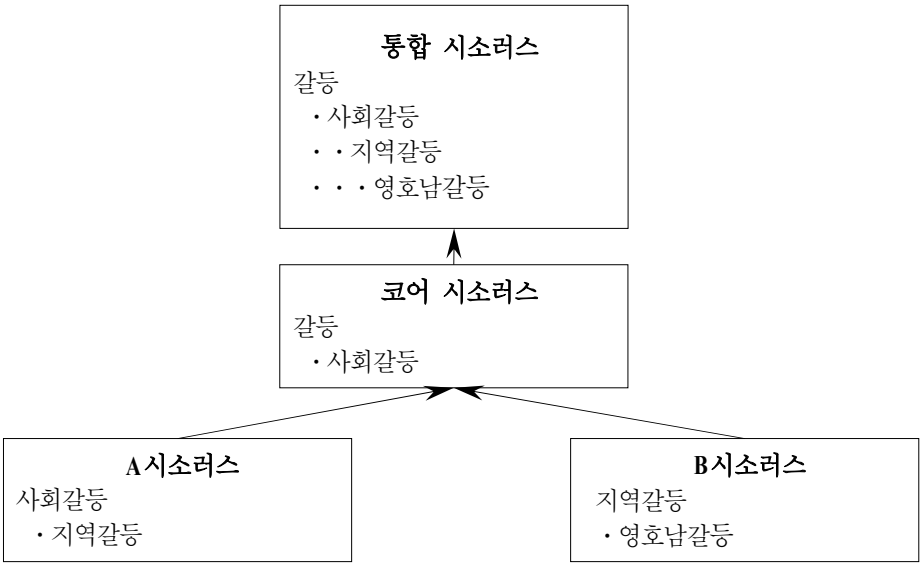
〈그림 5〉에서 나타낸 바와 같이 뉴스 코어 시소러스의 갈등의 하위어로 사회갈등이 있고, A시소러스의 사회갈등은 하위어로 지역갈등이 있으며, B시소러스의 지역갈등 하위어로 영호남갈등이 있을 경우, 하나로 병합하여 사용할 수 있다.



〈그림 5〉 시소러스의 병합

나. 통합

기존의 여러 마이크로시소러스가 활용되고 있고, 대형 포털 등에서 하나로 통합하여 서비스하는 것을 상정한 경우이다. 그 개념도를 〈그림 6〉에 나타낸다.



〈그림 6〉 시소러스의 통합

뉴스 코어 시소러스에서 더미텀으로 모으지 않으면, 각각 분리되어 있는 A시소러스의 환경보호와 B시소러스의 자연보호를 한 곳으로 모으지 못하는 경우가 발생할 수 있다. 뉴스 코어 시소러스의 더미텀을 통합하는 최상위 용어를 사용하여 흩어져 있는 것을 한 곳으로 모으고, 계층의 깊이가 달라도 통합해서 사용할 수 있는 방안이다.

3. 단독 사용

매크로시소러스와 마이크로시소러스를 전제로 구축하였으나, 병합이나 통합을 염두에 두지 않거나 좁은 영역을 차지하는 신문에서 단독으로 사용하는 방안이다. 이것은 넓은 주제 분야를 가진 주간이나 격주간 등과 같이 정보량이 적은 경우에 사용하는 경우이다.

4. 카테고리의 활용

마이크로시소러스의 카테고리를 주제분류로 사용할 수 있다. 더욱 확장하여 사용하는 하나의 방법으로 각각의 마이크로시소러스의 카테고리를 「a 기관단체」, 「b 사람」, 「c 행정/정책」, 「d 행사」,

「f 사건사고」, 「g 시설」, 「h 행위/활동」, 「j 시간」, 「k 공간」의 패싯으로 분류할 수 있다. 이것은 기사자료 표준분류표(1991)의 것을 확장한 것이다.[41] 영문자 소문자로 기호화한 것은 일반 카테고리와 구분하기 위함이고 a 이외의 모음을 쓰지 않은 것은 확장성과 식별성을 고려한 것이다.

5. 공유 및 상호운용성의 확보 방안

구축된 뉴스 코어 시소러스가 마이크로시소러스로 확장되고 매크로시소러스로서 기능하여, 마이크로시소러스와 병합이나 통합하여 활용되기 위해서는 뉴스 커뮤니티에서 실제로 공유되고 상호운용성을 확보할 필요가 있다. 이를 위한 방안으로서는 예를 들어 뉴욕타임스와 같이 링크드 데이터로서 자신의 시소러스를 공개하는 방안이다. 이것은 DBpedia나 Freebase, Geonames가 대응하는 토픽과 상호링크하고 있다. 그 의도는, 뉴욕타임스에 의해 유지되고 있는 콘텐트 아카이브로 사람들을 유도하기 위한 것으로서 높은 자유도로 라이선스된 데이터를 사용하는 것이다.[42]

또한, NewsML과 같은 뉴스 커뮤니티에서 표준화된 포맷으로 공유하는 것이다. NewsML이란 IPTC(International Press Telecommunications Council, 국제언론통신평의회)가 발표한 국제 표준 뉴스 포맷이다. NewsML은 원래 뉴스 교환을 위한 표준 포맷을 정하는 목적으로 설계되었으나 이제는 유통뿐만 아니라 아카이브 구축, 뉴스의 작성, 편집, 관리, 출판의 전 영역을 지원할 수 있는 것으로 인정받고 있다.[43] 이 NewsML에는 1,397항목의 뉴스 코드(뉴스 분류체계)가 존재하고 있다. 국내에서는 뉴스ML 시소러스가 이미 종합지 위주의 표준화된 시소러스가 이용되고 있으며, 같은 방식으로 뉴스 코어 시소러스를 활용할 수 있을 것이다.

한편, SKOS로 기 구축된 뉴스 코어 시소러스를 메타데이터 레지스트리로서 공개하여 상호운용성을 확보할 수 있다. 메타데이터 레지스트리는 메타데이터의 상호운용성을 확보하기 위한 목적으로 고안된 것이다. 메타데이터의 등록과 인증을 통하여 표준화된 메타데이터를 유지 관리하며, 메타데이터의 명세와 의미의 공유를 통해 메타데이터 집합 또는 메타데이터 요소 간의 호환성을 유지시킨다.[44] 이 메타데이터 레지스트리를 이용하여 상호운용성을 확보하는 방안이다. 메타데이터 레지스트리는 국제표준협회 산하의 32 하부위원회(ISO/IEC JTC1 SC32)에서 개발한 국제표준이다.

41) 한국언론연구원 · 한국조사기자회, 기사자료 표준분류표(서울 : 한국언론연구원 · 한국조사기자회, 1991), p.4.
42) Tom Heath and Christian Bizer, Linked Data: Evolving the Web into a Global Data Space, 2011. [cited 2013. 6. 20].
43) 한국언론재단, NewsML의 이해(서울 : 한국언론재단, 2007). p.12.
44) 고영만, "메타데이터 표준화와 메타데이터 레지스트리," 국회도서관보, 제42권, 제11호(2005. 11), p.20.

V. 결 론

우리나라의 신문은 2012년을 기준으로 일반일간 239, 일반주간 953, 인터넷신문 3,918, 외국일간 9, 통신 14, 특수일간 113, 특수주간 1615개 등 총 6,861개가 있다. 신문의 양도 많고, 발행주기에 따라 일간, 주간 등으로 나눌 수 있고, 종합지, 경제지, 스포츠지 등의 분야별로 나눌 수 있다. 이와 같은 다양한 신문의 뉴스 정보를 효율적으로 관리하기 위해서는 시소러스가 필요하다. 그러나 통합된 하나의 시소러스를 유지하면, 유지하는 노력도 많이 소요되고, 시소러스 용어의 수에 따라 어떤 신문은 너무 과다하고 어떤 신문은 너무 적을 수 있다. 뉴스의 모든 영역을 커버할 수 있는 뉴스 코어 시소러스를 구축하고, 이것을 매크로시소러스로 하고 다양한 분야의 시소러스를 마이크로시소러스로 하여 관리하는 방안을 제시하였다.

본 연구의 결과를 종합하면 다음과 같다.

첫째, 뉴스 관련 분류표 및 시소러스, Daum 디렉토리, 문화체육관광부 및 시도에 등록된 정기간행물 중 일간신문, 주간신문, 특수신문, 인터넷신문 등을 분석하여 매크로시소러스와 마이크로시소러스의 구성 체계를 제시하였다. 구성 체계는 뉴스 코어 시소러스를 매크로시소러스로 하고 일간종합신문을 대상으로 하는 신문 종합, 일간경제신문을 대상으로 하는 경제 종합, 스포츠신문을 대상으로 하는 스포츠/연예, 그리고 사회, 문화, 자연, 국제 영역의 각각을 대상으로 하는 사회 종합, 문화 종합, 자연 종합, 국제 종합 등으로 마이크로시소러스를 구성할 수 있다. 신문 종합을 제외한 각각의 마이크로시소러스는 코어 시소러스와 해당 마이크로시소러스를 매크로시소러스로 하는 각각의 세부 마이크로시소러스를 두는 것으로 구성하였다.

둘째, 각종 뉴스 관련 분류표의 중간 분야(강)와 시소러스의 최상위 및 고립어를 수집하고 분석하여 뉴스 코어 시소러스를 구축하였다. 완성된 시소러스는 우선어 2,012어, 비우선어 74어, 더미텀은 117어 및 대분류 15항목, 중분류 87항목이었다.

셋째, 뉴스의 주제 영역을 마이크로시소러스로 구성하는 분류는 분류표로도 활용할 수 있도록 구성하였다. 더욱 확장하여 사용하는 하나의 방법으로 각각의 카테고리를 「a 기관단체」, 「b 사람」, 「c 행정/정책」, 「d 행사」, 「f 사건사고」, 「g 시설」, 「h 행위/활동」, 「j 시간」, 「k 공간」의 패싯으로 분류할 수 있다.

셋째, 더미텀을 두어 관련 용어를 한 곳으로 모으고, 향후 매크로시소러스와 마이크로시소러스의 병합 및 통합에 활용할 수 있게 하였다.

넷째, 구축된 뉴스 코어 시소러스를 기존의 시소러스와 병합 또는 통합, 정보량이 적은 소규모의 특수 신문을 저비용으로 시소러스를 확장하여 활용할 수 있는 방안을 제시하였다.

다섯째, 뉴스 코어 시소러스를 SKOS로 작성하여 활용도를 높였다.
여섯째, 공유 및 상호운용성 확보 방안을 제시하였다.

뉴스 코어 시소러스를 구축하고 활용하는 이점으로는 다음과 같은 것을 들 수 있다.
첫째, 뉴스 코어 시소러스와 함께 필요한 모듈만을 선택하여 사용할 수 있고, 포털 사이트나 뉴스 아카이브에서는 병합하거나 통합하여 사용할 수 있다.
둘째, SKOS를 활용하여 공개하면, 소규모의 신문사의 경우 고비용의 툴이나 용어 관리 부담이 없이, 저비용으로 시소러스를 활용할 수 있다.
셋째, 여러 신문사가 함께 뉴스 코어 시소러스를 사용하면 상호운용성을 확보할 수 있다.

자동 병합이나 통합은 추후 연구과제이다. 자동 병합이나 통합의 경우 동형이의어와 띄어쓰기 등에 의해 수작업이 불가피할 것으로 사료된다. 이들의 연구가 더욱 진행되면 타 영역에서의 시소러스의 활용에도 도움이 될 수 있을 것이다.

참고문헌

고영만. “메타데이터 표준화와 메타데이터 레지스트리.” 국회도서관보, 제42권, 제11호(2005. 11), pp.18-26.

문화체육관광부(정기간행물등록현황) 홈페이지 〈http://www.mcst.go.kr/web/s_data/deptData/deptDataView.jsp?pSeq=81〉 [인용 2013. 5. 3].

장명희. 한국어 매크로 시소러스(Korean macro-thesaurus)의 작성법 연구. 석사학위논문, 숙명여자대학교 대학원 국어국문학과, 2001.

최석두. “한글 매크로시소러스 구축의 실제.” 한국정보관리학회 학술대회 논문집, 제5호(1998), pp.223-226.

한국경제신문사. 경제신문시소러스. 서울 : 한국경제신문사, 1993.

한국언론연구원. 신문기사 종합시소러스. 서울 : 한국언론연구원, 1993.

한국언론연구원 · 한국조사기자회. 기사자료 표준분류표. 서울 : 한국언론연구원 · 한국조사기자회, 1991.

한국언론재단. 뉴스ML 시소러스(상)(하). 서울 : 한국언론재단, 2008.

한국언론재단. NewsML의 이해. 서울 : 한국언론재단, 2007.

廣木守雄, 服部信司. ニュースシソーラス. 東京 : 日外アソシエーツ, 2004.

朝日新聞社ニューメディア本部. 朝日記事データベース分類の手引き. 東京 : 朝日新聞社ニューメディア本部, 1989.

Aitchison, J. Gilchrist, A. and Bawden D. *Thesaurus construction and use: a practical manual*. 4th ed. London : Aslib imi, 2000.

Aranda, María Mitre. Tesqual: *A Microthesaurus for Use in Quality Management in European Higher Education*, 2010. 〈http://www.intechopen.com/books/quality-management-nd-six-sigma/-tesqual-a-microthesaurus-for-use-in-quality-management-in-european-higher-education-〉 [cited 2013. 6. 20].

Breuker, Joost. Valente, Andre. and Winkels Radboud. "Use and Reuse of Legal Ontologies in Knowledge Engineering and Information Management." In: *Law and the Semantic Web*, edited by Benjamins et al. Berlin Heidelberg : Springer-Verlag, 2005. pp.36-64.

Daum 디렉토리 〈http://directory.search.daum.net/site_list.daum〉 [인용 2013. 5. 15].

EUROVOC Home page. 〈http://eurovoc.europa.eu/drupal/〉 [cited 2013. 5. 1].

Government of Canada Core Subject Thesaurus. 〈http://www.thesaurus.gc.ca/default.asp?lang=En&n=0073D232-1〉 [cited 2013. 5. 15].

Heath, Tom and Bizer, Christian. *Linked Data: Evolving the Web into a Global Data Space*, 2011. 〈http://linkeddatabook.com/book.〉 [cited 2013. 6. 20].

IPTC Home page. 〈http://cv.iptc.org/newscodes/subjectcode/〉 [cited 2013. 5. 15].

ISO 2788. 시소러스 개발 지침. 최석두. 정동열 공역. 서울 : 문헌정보처리연구회, 1994.

Lancaster, F. Wilfrid. 情報システムのための 構築と利用. 松村多美子. 鈴木 祐滋 譯. 東京 : 情報科學技術協會, 1989.

Lancaster, F. Wilfrid. *Vocabulary Control for Information Retrieval*. 2rd ed. Arlington : Information Resources Press, 1986.

Nikkei Telecom Home page. 〈http://www.nikkeitel.com/〉 [cited 2013. 5.1].

OECD Development Centre. *Macrothesaurus for Information Processing in the Field of Economic and Social Development*. 5th Ed. 1998. 〈http://www.oecd-ilibrary.org/development/macrothesaurus-for-information-processing-in-the-field-of-economic-and-social-development_9789264162990-en〉 [cited 2013. 6. 20].

Soergel, Dagobert. *Indexing Languages and Thesauri: Construction and Maintenance*. Los Angeles : Melville Publishing Company, 1974.

The Washington Post. *The Washington Post Thesaurus*. Washington : The Washington company, 1986.

W3C Home page. 〈www.w3.org/2004/02/skos/core/owl-dl/skos-core-owl-dl.owl〉 [cited 2013. 5. 15].

Willpower Information. 〈http://www.willpowerinfo.co.uk/glossary.htm〉 [cited 2013. 5. 15].

국한문 참고문헌의 영어 표기
(English translation / Romanization of references originally written in Korean)

Choi, Suk-Doo. "Practical Construction of Hangul Macro-Thesaurus." 5th Proceedings of the Korean society of Information Management, 5(1998), pp.223-226.

Daum directory Home page. 〈http://directory.search.daum.net/site_list.daum〉 [cited 2013. 5. 15].

Hiroki, Morio, Hatori, Shinji. *News Thesaurus*. Tokyo: Nichigai Associates (publisher), 2004.

Jang Myeong-Hee. *A Study of a Method of Drawing up the Korean Macro-thesaurus*. M.A. thesis, Department of Korean Language and Literature Graduate School Sookmyung Women's university, 2001.

Korea Press Foundation. *NewsML Thesaurus(I)(II)*. Seoul : Korea Press Foundation, 2008.

Korea Press Foundation. *Understanding NewsML*. Seoul : Korea Press Foundation, 2007.

Korea Press Institution. *News Articles Thesaurus*. Seoul : Korea Press Institution, 1993.

Ko, Young-Man. "Metadata Standard and Metadata Registry." *National Assembly Library Review*, Vol.42, No.11(2005. 11), pp.18-26.

Ministry of Culture Sports and Tourism Home page. 〈http://www.mcst.go.kr/web/s_data/deptData/deptDataView.jsp?pSeq=81〉 [cited 2013. 5. 3].

The Korea Economic Daily. Economic News Thesaurus. Seoul : The Korea Economic Daily, 1993.