신문 시소러스 개발의 이론과 실제

정 영 미*

목 차

1. 서 론

효과적인 신문기사 검색시스템에 관한 요구는 이미 오래전부터 있어 왔으나 특히 최근 들어 신문기사 데이타베이스의 구축과 함께 그 필요성이 더욱 증대되고 있다. 신문기사 데이타베이스는 다른 유형의 데이타베이스와는 달리 대량의 기사 전문을 수록하고 있으며 이로부터 정보이용자가 원하는 기사를 그대로 검색할 수 있다는 점에서 그 가치가 더욱 크다.

그러나 신문기사는 양적인 면에서 다른 정보자료와는 비교할 수 없을 정도로 엄

* 연세대학교 문헌정보학과 교수

청나기 때문에 그 데이타베이스로부터 정보이용자가 정말 원하는 정보를 검색하기가 그만큼 어려울 수 밖에 없다. 검색의 효율성은 결국 색인이 얼마나 정확하게 되어 있는가에 의해 결정되므로 효과적인 정보검색시스템의 구축을 위해서는 무엇보다도 효과적인 색인시스템의 개발이 선행되어야 할 것이다.

신문기사의 색인은 색인대상이 되는 기사량을 고려할 때 일차적으로는 수작업 방식에 의존하기보다는 자동색인 방식을 사용하는 것이 효율적일 것으로 보인다. 그러나 현재의 자연언어 처리 기술로는 색인전문가가 작성한 색인에 필적하는 색인을 컴퓨터가 생산하기란 거의 불가능하다. 특히 자동색인 방식에 의해 색인할 경우에는 용어의 통제가 어려우므로 유사한 개념을 나타내는 다양한 용어가 색인어로 선택되는 문제가 발생한다. 특히 신문기사의 경우에는 주제의 폭이 넓고 기사 작성자가 다수이므로 다양한 어휘가 사용되어 표현의 일관성을 기대하기가 더욱 어렵다. 이러한 문제는 색인전문가에 의한 색인에서도 용어의 통제가 이루어지지 않는 경우에는 마찬가지로 발생한다. 따라서 색인전문가가 색인작업을 수행하되 일관성있고 효과적인 색인을 생산함으로써 검색의 효율성을 제고할 수 있는 방안을 모색하는 것이 우선적으로 해결해야 할 과제일 것이다.

본 연구에서는 경제신문용 시소러스 개발을 목표로 하여 우리말 신문기사 색인어의 선정 원칙을 수립하고, 시소러스 개발 과정에서 검토해야 할 문제점들을 예와 함께 기술하고자 한다. 또한 이 연구 결과 구축된 시소러스를 신문기사의 색인과 검색에 활용하는 방안을 제시하고자 한다.

2. 신문기사 검색시스템의 유형과 평가

2.1 자연언어 색인/자연언어 탐색

색인어와 탐색어로 모두 자연언어를 사용하는 시스템은 신문기사 데이타베이스를 검색하는 시스템에서 고려해볼 만 하다. 왜냐하면 자동색인의 대상이 되는 전문 데이타베이스가 구축되어 있고 또한 탐색자는 시소러스와 같은 통제언어에 익숙하지 않은 최종이용자이기 때문이다. 그러나 앞에서도 언급했듯이 자동색인 방식에

의해 자연어 키워드(자유키워드)를 색인어로 선택할 경우 동일한 개념을 다른 키워드로 표현할 수가 있으므로 탐색자가 정보요구와 관련된 모든 용어를 탐색어로 사용하지 않는 한 검색효율이 떨어지게 된다. 또한 전문 데이타베이스에 적용할 수 있는 자동색인 기법이 아직은 통계적 기법이나 간단한 구문분석기법에 한정되어 있으므로 해당 기사의 주제어가 아닌 키워드를 색인어로 선정하는 문제가 발생하게 되며 이것 또한 검색효율을 저하시키는 원인이 된다.

수작업에 의한 색인시스템에서도 시소러스가 없이 색인하는 경우에는 자동색인에서와 같이 색인어 선정에 있어서 일관성이 결여되는 문제가 발생한다. 실제로 KETEL에서 지난 1년간 축적한 색인어휘를 조사해 본 결과 다음예에서와 같이 일관성없는 색인어들이 선정되어 있는 경우가 허다했으며, 이것은 통제어휘집이 없이 색인할 때 발생하는 전형적인 문제일 것으로 보인다(괄호안은 빈도임).

유엔안전보장이사회(34): UN안전보장이사회(2): 안전보장이사회(10)
추경예산(34): 추가경정예산(4): 85년 추경예산(1): 87년 추경예산(1)
GNP(58): 국민총생산(43)
코메콘(15): COMECON(7): 코메콘개편(3): 코메콘개혁(1)
국제유가(20): 국제원유가(29): 국제원유가격(22): 원유가격(38)
연두기자회견(52): 대통령연두회견(1): 연두회견(1)
VAN(31): 부가가치통신망(19)
테니스볼(3): 테니스공(1)

자연언어 색인에서는 이와 같이 주제어만을 색인어로 선택한다고 해도 동일한 주제를 다루는 기사들이 여러 개의 색인어 아래 분산되므로 탐색시 탐색어로 선택한 용어 아래 색인된 기사만이 검색되고 나머지 관련기사는 검색되지 않는 결과를 초래한다. 결국 탐색자의 입장에서는 탐색의 자유를 누리는 대신 자신의 정보요구에 대해 부적합한 문헌의 검색과 동시에 적합한 문헌의 누락을 감수하지 않을 수 없다. 따라서 자연언어에 의한 색인 및 탐색은 통제언어에 의한 색인과 탐색이 가능한 시스템에서 이용자 편의를 고려하여 함께 병행하는 것이 바람직하다.

이러한 유형의 시스템에서 검색효율을 높일 목적으로 탐색시 동의어사전과 유사한 탐색용 시소러스를 사용하여 탐색어를 추가할 수 있다. 이 탐색용 시소러스는 동

의어, 유사동의어, 약어 등을 수록하고 있어 탐색자가 현재 사용하고자 하는 탐색어와 개념적으로 유사한 다른 탐색어를 추가하는 데 도움을 준다. 예를 들어 '유엔안전보장이사회'를 탐색어로 사용하고자 할 때 'UN안전보장이사회'와 '안전보장이사회'를 탐색어로 추가할 수 있을 것이다. 이러한 탐색용 시소러스는 책자 형태보다는 컴퓨터 화일로 구축하여 탐색시 이용자가 온라인으로 참조하도록 하면 자연언어가 갖는 특정성을 유지하면서 동시에 포괄적인 탐색이 가능해진다.

2. 2 통제언어 색인/통제언어 탐색

색인과 탐색시 모두 통제언어를 사용하는 시스템은 시소러스에 수록되어 있는 디스크립터(통제키워드)를 색인어로 선정하고 탐색시에도 같은 용어를 탐색어로 사용함으로써 검색효율을 극대화시킬 수 있다. 이와 같이 색인과 탐색시 통제어휘에 의존하는 시스템은 첫째, 특정한 개념은 항상 같은 용어에 의해 색인되므로 검색효율이 높으며, 둘째, 시소러스에 나타나 있는 용어간의 어의적 관계를 이용하여 일차적으로 선택한 탐색어 외에 다른 용어를 탐색어로 추가하여 탐색영역을 확장할 수 있다는 장점을 갖는다.

예를 들어 다음과 같은 시소러스 엔트리가 있으며 '교포'에 관한 기사를 원할 경우, 탐색자가 '해외교포', '해외동포' 등의 용어를 탐색어로 먼저 생각했다면 '해외교포 USE 교포'와 같은 참조 엔트리를 이용하여 먼저 생각한 비디스크립터 대신 디스크립터인 '교포'를 탐색어로 사용하게 된다. 또한 '소련교포', '재일교포', '재미교포' 등 보다 특정한 개념의 용어를 탐색어로 추가함으로써 탐색을 확장할 수 있다.

교포	
UF	해외교포
	해외동포
	재외국민
	교민
NT	소련교포

재미교포
재일교포
중국교포

반면에 색인어와 탐색어를 통제하는 시스템의 단점으로는 첫째, 자연언어 시스템에 비해 용어의 특정성이 덜하므로 주제의 구체적인 표현이 어렵다는 점, 둘째, 통제어휘에 익숙하지 않은 최종이용자에게는 불편한 시스템이라는 점, 셋째, 주제분야의 어휘 변화를 반영하도록 시소러스의 계속적인 갱신이 필요하다는 점 등이 있다. 그러나 위와 같은 단점들은 신문기사 색인 및 검색시스템에서 다음과 같이 극복할 수 있다.

먼저 용어의 특정성 문제는 신문기사의 경우 그다지 심각하지 않은 것으로 보여지는데, 그 이유는 신문기사는 그 특성상 복합어의 사용이 빈번하고 따라서 탐색자의 편의를 위하여 사용빈도가 높은 복합어(예: 수도권인구집중)는 그대로 디스크립터로 선정할 수 있기 때문이다. 또한 일시적인 위원회(예: 경제난국극복위원회)라든가 지역별 상품거래소(예: 뉴욕상품거래소, 시카고상품거래소) 등과 같이 일시적이거나 일일이 열거하기 힘든 용어는 자유키워드로 색인하도록 하여 높은 특정성을 유지할 수 있다. 또한 디스크립터로 선정하기에 특정성이 너무 높은 개념의 용어나 특정어로 처리하는 것이 나은 용어는 다음 예에서와 같이 상위개념어를 색인어로 선택하도록 지시하거나 두 개의 디스크립터로 색인하도록 지시하여 색인작업시 이 개념이 누락되는 일이 없도록 한다.

경기선행지수상승률 USE 경기선행지수 + 상승률
전세금인상 USE 전세금 + 인상(또는 가격인상)
정국불안 USE 정국
신규인력채용계획 USE 신규채용

두번째의 문제를 해결하는 방안은 시소러스를 온라인으로 탐색할 수 있도록 하여 탐색자가 자신이 생각하고 있는 용어가 탐색어로 적합한지를 확인하고 적합하지 않은 경우 적합한 다른 용어를 선택하는 데 도움을 주도록 하는 것이다. 세번째

의 어휘변화의 문제는 신문기사의 경우 특히 심각하게 보이는데 어휘변화에 대처하는 방안은 우선 일시적이거나 지나치게 전문적인 용어는 시소러스에 수록하는 대신 색인자가 필요에 따라 자유키워드로 선택하게 하여 시소러스 내의 어휘변화를 최소화하는 것이다. 이러한 용어들은 별도의 전거화일에 수록하여 관리하는 것이 바람직하다.

2. 3 통제언어 색인/자연언어 탐색

색인시에는 통제된 색인어인 디스크립터를 부여하고 탐색시에는 자연어 키워드를 사용하는 시스템은 탐색자 입장에서 볼 때에 높은 검색효율을 유지하면서 동시에 사용하기 편리한 시스템으로서 비교적 이상적인 시스템이라고 볼 수 있다. 이러한 시스템에서 색인자는 시소러스에 근거하여 색인하지만 탐색자는 자연어 탐색어를 입력하게 되는데, 입력된 탐색어는 컴퓨터에 내장된 탐색용 시소러스화일에 의하여 해당되는 디스크립터로 변환된 다음 시스템의 탐색어로 사용된다.

탐색용 시소러스는 일반적인 시소러스의 동등관계(USE/UF 참조)를 나타내는 엔트리(예: 안전보장이사회 USE 유엔안전보장이사회)들이 모여서 구성된 것과 같으며, 만일 탐색자가 '안전보장이사회'를 탐색어로 입력할 경우 이 용어는 디스크립터인 '유엔안전보장이사회'로 자동변환되어 탐색어로 사용된다.

이때 탐색자는 통제어휘에 익숙할 필요가 없이 자연언어를 사용할 수 있는 장점이 있는 반면, 용어의 자동변환을 위해 상당한 크기의 도입어휘를 컴퓨터에 온라인으로 소장하는 데 따르는 어려움을 감안해야 한다. 또한 동음이의어나 다의어가 존재할 뿐 아니라 자연언어와 통제언어간의 용어의 특정성의 차이에서 오는 어려움 때문에 완벽한 변환을 기대하기 힘들다. 그러나 현대의 정보검색시스템은 이용자편의시스템(user-friendly systm)을 지향하고 있으므로 이러한 유형의 시스템에 대한 보다 적극적인 검토가 필요할 것이다.

2. 4 시소러스를 이용한 신문기사 검색시스템 모형

앞에서 언급한 신문기사 검색시스템의 유형을 평가한 결과 현 단계의 신문기사

검색시스템은 일차적으로 색인자는 시소러스를 이용하여 색인하고, 탐색자는 시소러스를 온라인으로 탐색하여 적절한 탐색어를 선정하는 '통제언어 색인/통제언어 탐색' 시스템이 되어야 할 것으로 보인다.

다음 단계로는 방대한 도입어휘를 포함하는 동의어 사전화일이나 이보다 훨씬 정교한 탐색용 시소러스화일을 이용하여 자연어 키워드를 디스크립터로 자동변환할 수 있는 시스템을 개발함으로써 '통제언어 색인/자연언어 탐색'이 가능하도록 하는 것이다.

한 단계 더 나아가서 지금보다 효과적인 신문기사 자동색인 기법이 개발된다면 '자연언어 색인/자연언어 탐색' 방식을 도입하여 '통제언어 색인/통제언어 탐색' 방식과 병행하여 사용하는 융통성있는 신문기사 검색시스템으로 발전시킬 수 있을 것이다. 이렇게 되면 탐색자가 자연어 키워드와 디스크립터를 구별없이 사용할 수 있으므로 시스템의 편의성과 검색효율을 동시에 증대시킬 수 있을 것으로 기대된다.

다음 절에서 언급할 일본 日刊工業新聞社의 기사정보 데이타베이스에는 각 신문기사에 대해 14개 이내의 디스크립터(통제어 키워드)와 7개 이내의 자연어 키워드가 수록되어 있어서 통제언어에 의한 탐색과 자연언어에 의한 탐색을 병용할 수 있도록 하고 있다. 또한 1991년부터는 시소러스 내의 디스크립터를 색인어로 입력하면 자동으로 이에 해당하는 비디스크립터가 색인어로 추가되므로 비디스크립터에 의한 탐색도 가능하도록 하였다. 즉 「日刊工業シソ-ラス」에 'LED USE 발광다이오드'라는 엔트리가 있다면 '발광다이오드'를 색인어로 부여할 때 자동적으로 'LED'가 추가되도록 한 것이다. 이것은 다시 말해 'LED'라는 비디스크립터를 탐색어로 입력하더라도 '발광다이오드' 아래 색인된 기사가 모두 검색됨을 의미한다.

3. 신문기사 시소러스의 사례 분석

3.1 한국언론연구원의 「신문기사 종합시소러스」[1]

이 시소러스는 한국언론연구원이 운영하는 언론종합정보은행인 KINDS를 위한

1) 한국언론연구원, 「신문기사 종합시소러스 :Kinds 신문기사 정보 검색용어집」 1993.

색인 및 검색도구로서 개발되었으며 국내 종합일간지의 기사색인을 위해 사용할 수 있다. 「신문기사 종합시소러스」의 주요한 특징은 다음과 같다.

(1) 시소러스의 체계는 자모순 시소러스(본체)와 분류별 계층색인, 분류별 자모순 색인으로 구성된다.

(2) 시소러스에 수록된 색인어에는 일반주제명 키워드, 기관단체명 키워드, 지역 키워드가 있으며, 인명 키워드와 기업명 키워드는 취급하지 않는다.

(3) 시소러스에 수록된 용어의 총수는 10,705개로서 이 가운데 디스크립터(우선어)는 7,179개, 비디스크립터(비우선어)는 3,526개이며, 특정한 개념은 상위개념어를 사용하도록 참조해 줌으로써 디스크립터의 수를 통제하였다.

(4) 용어간의 관계로는 동등관계, 계층관계, 연관관계를 설정하였으며, 스코프노트(범위주기)에 용어의 사용범위에 대한 주기를 기술하였다.

3. 2 日本經濟新聞社의 「日經シソ-ラス」[2]

「日經シソ-ラス」는 日本經濟新聞社의 기사정보데이타베이스에 색인어로 부여되는 10종류의 키워드 가운데 일반주제를 나타내는 3종류의 키워드(품목, 업계, 항목)와 지역 키워드를 통제키워드로 수록하고 있다. 회사명, 단체명, 인명은 시소러스에 수록하지 않고 별도로 취급하며, 특수한 개념이나 새로운 사상을 표현하는 용어, 전문용어나 시사용어 등은 자유키워드로 색인하도록 지시하고 있다. 이 시소러스는 계층색인 형태의 분야별 시소러스(본체)와 50음순 색인으로 구성되며, 수록된 디스크립터의 수는 15,434개이다. 50음순 색인에는 약 740개의 동의어가 참조표시와 함께 수록되어 있다. 분야별 시소러스에는 용어간의 계층관계가 나타나 있으며 연관관계는 설정되어 있지 않다.

3. 3 中日新聞本社의 「ニユ-ス・シソ-ラス」[3]

20여년간에 걸쳐 신문으로부터 수집한 용어를 바탕으로 하여 작성한 일본 中日

2) 日本經濟新聞社, 「日經シソ-ラス」 1990.

3) 中日新聞本社, 「ニユ-ス・シソ-ラス:新聞情報管理のための用語集」 1990.

新聞社의 시소러스(1990년판)는 50음순 시소러스 본체와 분야별 색인으로 구성되어 있다. 수록된 용어수는 모두 9,404개로서 7,473개의 디스크립터와 1,931개의 비디스크립터를 포함하며, 용어간의 동등관계, 계층관계, 연관관계가 표시되어 있다. 디스크립터는 일반주제 키워드 외에 기관/단체명을 포함하며 인명과 기업명은 포함하지 않는다.

3.4 日刊工業新聞社의 「日刊工業シソ-ラス」[4]

일본 日刊工業新聞社의 시소러스는 日刊工業新聞에 게재된 기사 가운데 특히 중요도가 높은 제품명, 새로운 기술 및 기업동향과 관련된 기사를 수록하고 있는 기사정보 데이타베이스를 위해 작성된 것이다. 이 시소러스(제3판)는 디스크립터 12,450개와 비디스크립터 553개를 수록하고 있으며, 시소러스 본체와 주제범주별 분류표로 구성되어 있다. 이 시소러스에는 용어간의 동등관계와 계층관계가 설정되어 있다. 기관/단체명은 별도로 「日刊工業 企業·團體コ-ドブック」에 코드번호와 함께 수록하여 전거화일을 구성하고 있다.

4. 신문기사 색인시 우리말 키워드의 선정 원칙

4.1 디스크립터의 선정원칙

디스크립터 선정을 위해 적용한 일반적인 기준은 다음과 같다.

(1) 용어의 시스템내에서의 이용빈도를 고려한다.

(2) 용어의 상대적 출현빈도를 고려한다.

(3) 이미 선정된 용어와의 관련성을 고려한다.

(4) 탐색시 사용되리라고 예상되는 용어를 선택한다.

시스템내에서의 이용빈도는 특히 특정성이 높은 복합명사의 선정시 고려하며, 용어의 상대적 출현빈도는 동의어, 약어/완전어, 외래어/번역어 등 동등관계에 있는

4) 日刊工業新聞社, 「日刊工業シソ-ラス」 제3판, 1990.

용어들 중에서 디스크립터를 선정할 때 참고한다. 즉 빈도가 높은 복합명사는 특정성이 아주 높더라도 디스크립터로 선정하거나(예: 중소기업고유업종) 또는 특정어로 처리하여 다른 디스크립터들을 조합함으로써(예: 중소기업→정보화 중소기업+정보화) 그 개념을 색인할 수 있도록 한다. 또한 동등관계에 있는 용어들은 다른 원칙이 적용되지 않는 한 빈도가 더 높은 용어를 디스크립터로 선정한다. 각각의 특정한 경우에서의 색인어 선택에 관해서는 다음에 상세히 기술하였다.

4.2 복합명사의 처리

신문기사 색인에 있어서 가장 문제가 되는 것은 복합명사로 표현되는 복합적인 개념을 어떻게 색인하는가 하는 것이다. 복합명사 색인어의 문제는 구체적으로 다음의 몇가지 경우로 구분하여 생각할 수 있다.

(1) 여러 개의 키워드가 결합된 경우
예: 세계경제성장률; 경기선행지수상승률; 중소기업국제회의

(2) 국가명과 키워드가 결합된 경우
예: 미해군; 미국방부

(3) 국가명과 국가명이 결합되고 그 다음에 키워드가 오는 경우
예: 한소경제협력; 한베트남경제협력; 미일경제협력

(4) 지명과 키워드(기관/단체명)가 결합된 경우
예: 시카고상품거래소; 대구상공회의소; 뉴욕증권거래소

(5) 상품명/물품명과 다른 키워드가 결합된 경우
예: 신발수출; 수입쇠고기; 담배수입; 반도체시장개방

(6) 전시회, 위원회 등 일시적이며 특정한 명칭의 경우
예: 캐나다상품특별전시회; 소련민속전시회; 한국상품전시회;
엑스포92; 파리문구박람회; 경제난국극복위원회

(7) 일시를 나타내는 숫자와 키워드가 결합된 경우
예: 5. 8 부동산투기억제대책; 6. 10 민주항쟁계승국민대회;
7. 4 남북공동성명; 8. 30 증시대책; 500일 경제대책

4.2.1 복합명사의 처리 원칙

「ISO 2788」[5] 과 「Unesco 지침서」[6]에 따른 복합명사의 처리원칙은 다음과 같다.

(1) 복합어는 원칙적으로 단일개념으로 분해한다.

(2) 일상적으로 사용하는 복합어는 분해하지 않는다.

(3) 복합어를 구성하는 각 요소가 시소러스내 별개의 디스크립터로 나타나는 경우 분해한다.

(4) 구성요소가 하나의 복합어 속에서만 사용될 경우 분해하지 않는다.

(5) 복합어를 구성하는 단어의 의미가 원래의 의미와 다를 경우 분해하지 않는다.

(6) 특정한 분야에서 매우 빈번하게 출현하는 복합어로서 분해한 결과 이용자가 불편을 느낄 경우 분해하지 않는다.

(7) 고유명사를 포함하는 복합어는 분해하지 않는다.

4.2.2 표준원칙을 준용한 복합명사의 처리 방안

본 연구에서는 4.2.1의 원칙을 준용하여 4.2에서 구분한 문제들을 처리하였다. 개별적인 복합명사의 처리 결과는 시소러스에 나타나 있으므로 여기에서는 대표적인 경우에 한하여 처리 방안을 기술하고자 한다.

(1) 여러 개의 키워드가 결합된 복합명사의 경우:

복합어를 구성하는 각 요소가 별개의 디스크립터로 나타나는 경우에는 물론 분해하며, '인상, 상승률, 조사, 성장률' 등과 같이 특정한 주제분야에 관계없이 디스크립터와 결합하여 복합명사를 구성하는 용어들은 공통적 디스크립터로 선정하여 이러한 단어와 결합하는 수많은 복합명사들을 분해할 수 있도록 하였다.

예: 전세금인상 → 전세금 + 인상(또는 가격인상)

물가상승률 → 물가 + 상승률

5) International Organization for Standardization, *ISO 2788: Guidelines for the Establishment and Development of Monolingual Thesauri*, 2nd ed. Geneva: ISO, 1986.

6) *UNISIST Guidelines for the Establishment and Development of Mono- lingual Thesauri*, 2nd ed. Paris: Unesco, 1981.

경기선행지수상승률 → 경기선행지수 + 상승률
세계경제성장률 → 세계경제 + 성장률
중소기업정보화 → 중소기업 + 정보화
중소기업국제회의 → 중소기업 + 국제회의
육군감축 → 육군 + 감축

그러나 '임금인상, 가격인상, 요금인상' 등과 같이 매우 빈번하게 사용되어 하나의 복합어로 정착된 용어는 분해하지 않으며, 다른 디스크립터와 결합하여 더 특정한 개념을 나타낼 수 있는 복합명사도 분해하지 않는다. 또한 구성요소가 하나의 복합어 속에서만 사용되는 경우에는 분해하지 않는다.

예: 중소기업고유업종; 경제난극복

(2) '수입, 수출, 시장개방' 등과 같이 그 대상이 되는 상품/물품/산업 등이 다양한 경우 상품명/물품명/산업명과 이에 관련된 디스크립터를 결합한 형태의 복합명사로 색인한다. 이때 '수출, 수입'과 결합되어 형성된 색인어는 일일이 시소러스에 수록하지 않고 자유키워드로 색인하는 것을 원칙으로 한다.

예: 쇠고기수입; 담배수입; 마늘수입; 신발수출; 수입쇠고기;
유통업개방; 반도체시장개방; 자동차시장개방

그러나 '수입쇠고기, 유통시장개방' 등과 같이 빈번히 나타나는 복합어는 시소러스에 수록할 수 있다. 또한 '수입농산물, 원자재수입' 등과 같이 수입, 수출의 대상이 되는 것이 특정한 물품이 아니라 범주를 나타내는 용어인 경우에는 시소러스에 수록하는 것이 좋다. 이렇게 함으로써 특정한 개념을 표현하는 디스크립터가 없는 경우 자유키워드를 사용하지 않더라도 그 개념을 포함하는 상위개념어로 색인할 수 있다. 즉 '마늘수입'과 같은 특정한 개념을 색인하지 않는 경우 '농산물수입'을 색인어로 선택하면 된다.

(3) 국가명, 지명 등의 고유명사와 다른 키워드가 결합된 복합명사는 일반적으로

통용되는 형태를 색인어로 선택한다. 색인 당시 시소러스에 디스크립터로 등록되어 있지 않은 용어는 그 중요성에 따라 후에 시소러스에 추가하거나 스코프 노트에서 지시한 대로 자유키워드로 색인한다.

예: 미해군: 한소경제협력: 대구상공회의소: 뉴욕증권거래소: 한미관계

(4) 일시적이거나 한시적인 위원회, 전시회, 행사, 협상, 회의 등의 명칭은 원칙적으로 시소러스에 수록하지 않으므로 기사에 나타난 형태 그대로 자유키워드로 색인한다. 그러나 '우루과이라운드'와 같이 지속적으로 신문에 나타나는 용어는 시소러스에 수록하는 것이 편리하다. 이러한 용어는 시소러스 갱신시 신문에 더 이상 출현하지 않음이 확인되면 시소러스에서 삭제하여 자유키워드로 환원한다.

예: 캐나다상품특별전시회: 전국노동자대회: 한미쇠고기 협상:엑스포92

(5) 날짜와 키워드가 결합된 용어들은 매우 특정하고 일시적인 용어이므로 그 자체를 디스크립터로 등록할 필요는 없다. 날짜를 뺀 나머지 키워드에 해당하는 디스크립터로 이 개념을 색인한다.

예: 5.8 부동산투기대책 → 부동산투기
1993년 예산 → 1993년 + 예산

4. 3 고유명사의 처리

인명, 지명, 기업명, 기관명 등의 고유명사는 특히 식별어(identifier)[7] 라고 하여 색인자가 임의로 색인어로 부여할 수 있다. 일반적으로 식별어는 시소러스에 포함

7) *Thesaurus of ERIC Descriptors*에서는 식별어를 (1)지역명, 인명, 프로그램명 등의 고유명사와, (2)아직 시소러스에 수록되지 않은 개념들의 두 가지 유형으로 구분하여 주고 있으며, 이러한 식별어들은 ERIC Identifier Authority List에 수록하여 둔다. 'writing improvement', 'distance location' 등은 식별어에서 디스크립터가 된 용어들의 예이다. (J.E. Houston, ed., *Thesaurus of ERIC Descriptors*, 10th ed. Phoenix: Oryx Press, 1984.)

하지 않거나 제한된 수만을 포함한다. 시소러스에 포함하지 않는 경우에는 별도의 전거화일에 수록하여 항상 일관성있게 한 개념을 표현할 수 있도록 하는 것이 중요하다. 왜냐하면 기업명, 기관명, 지명 등은 다양한 형태로 사용되기 때문에 어떤 형태로 접근하든 관련된 기사를 검색할 수 있도록 하는 장치가 필요한 것이다.

시소러스 작성에 관한 국제표준인 ISO 2788이 식별어 선정에 관해 제시하고 있는 몇가지 기본 원칙은 다음과 같다.

(1) 국제적 기관과 하나 이상의 언어로 문헌을 발행하는 지역적 기관은 이용자에게 가장 친숙한 형태로 기관명을 표기한다. 다른 언어로 접근할 가능성이 있는 경우에는 다른 형태로부터 참조를 내준다.

(2) 하나의 언어로 출판하거나 사업을 수행하는 국가적 기관과 지역적 기관은 번역하지 않은 형태의 명칭을 선택한다. 번역된 명칭이 있는 경우에는 이로부터 참조를 내준다.

(3) 개인명은 원칙적으로 원래의 형태로 기록한다. 단, 특정한 지역에서 사용되는 명칭이 시소러스 이용자에게 더 잘 알려져 있는 경우 이를 선택하고 원래의 이름으로부터 참조를 낸다.

고유명사 가운데 지명과 국명은 식별어로 취급할 수 있으나 대부분의 (시소러스에서 통제키워드로 처리한다. 中日新聞社의 시소러스에서는 '國際' 분야에서 '日本' 아래 국내지명을, '國名' 아래 외국지명과 국명을 수록하고 있다. 한국언론연구원의 시소러스에서도 마찬가지로 국내지역과 국외지역으로 나누어 수록하고 있다. 「日經シソ-ラス」에서는 '地域'이라는 분야 아래 국내지역과 해외지역으로 구분하여 지명을 수록하고 있다.

본 연구에서는 국명과 산, 바다, 강, 호수 등의 지명은 시소러스에 수록하되 신문기사에 빈번히 나타나는 것만을 수록하고 나머지는 자유키워드로 색인하도록 지시하였다.

ISO 2788은 둘 이상의 지명이 병행하여 사용되는 경우에는 이용자에게 친숙한 용어를 선택하며, 통칭과 공식적 명칭이 똑같이 친숙한 경우에는 공식적 명칭을 선택하도록 권하고 있다. 본 연구에서도 이 원칙에 따라 다음과 같은 엔트리를 작성하였다.

예: 벨기에 UF 벨지움
프랑스 UF 불란서
남미 UF 남아메리카

식별어 가운데 특히 문제가 되는 것은 기관/단체명이다. 기관/단체명은 다음의 세 가지 방법으로 처리할 수 있다. 첫째는 회사명/기업명과 함께 모든 기관/단체명을 전거화일에 수록하는 방법이다. 이 경우 기관/단체명 처리에 통일성이 부여되는 장점이 있으나 빈번히 사용되는 단체명을 시소러스에서 찾을 수 없는 불편함이 있다.

둘째는 공식적인 기관/단체명을 모두 시소러스에 수록하는 것인데, 이 경우 신문에 자주 출현하지 않는 많은 수의 단체명이 시소러스에 수록되므로 시소러스 어휘가 너무 커지는 문제점이 있다.

셋째는 신문기사에 출현하는 빈도가 높거나 중요하다고 판단되는 제한된 수의 기관/단체명만을 시소러스에 수록하고 나머지는 전거화일에 수록하는 방법이다. 이 방법은 중요한 기관/단체명은 시소러스에서 찾을 수 있으며 나머지 기관/단체명은 스코프노트를 통해 전거화일을 참고하도록 함으로써 시소러스 어휘의 수를 줄이면서 모든 기관/단체명을 통제할 수 있는 가장 효과적인 방법으로 생각된다. 현재 일본 中日新聞社의 시소러스와 한국언론연구원의 시소러스는 기관/단체명 가운데 잘 알려진 명칭만을 시소러스에 수록하고 있으며 수록되지 않은 명칭에 대한 전거화일은 따로 만들어 주지 않고 있다.

다음은 위의 세 가지 방법 가운데 본 연구에서와 같이 세번째 방법을 채택할 경우 생성되는 엔트리의 예이다.

예: 경제단체
SN 수록되지 않은 각 경제단체와 경제관련 연구기관은
전거화일을 보시오.
NT 경단협
대한상의
상공회의소

전경련
중소기업협동조합중앙회
한국경총
한국무역협회
한국생산성본부

4.4 약어의 처리

신문기사에서는 완전어보다는 약어를 선호하는 특성을 반영하여 다음과 같은 원칙을 설정하였다.

(1) 일반적으로 완전어 대신 약어가 잘 알려져 있는 경우에 약어를 디스크립터로 선택한다.

예: 동남아(←동남아시아)
CATV(←유선TV방송, 케이블TV방송)

(2) 기관/단체명의 경우 약어가 있으면 특별한 경우를 제외하고는 약어를 색인어로 선택한다.

예: 전경련(←전국경제인연합회)
전교조(←전국교직원노동조합)
ESCAP(←아시아태평양경제사회위원회)
IBRD(←세계은행)

(3) 영문 약어가 있으나 잘 알려져 있지 않거나 거의 사용되지 않는 경우에는 우리말 번역어를 선택한다.

예: 환매채(←RP)

국민순생산(←NNP)

신흥공업경제(←NIES)

4.5 외래어 처리

(1) 외국어로 표현된 용어가 현재 정착되어 있다면 외국어를 선택한다.

예: 팀스피리트훈련(←한미합동군사훈련)

(2) 선택된 외국어가 영문 약어가 아닌 경우에는 우리말로 표기한다.

예: 스왑거래(←스와프거래): 우루과이라운드(←UR)

(3) 영문 약어의 경우 우리말 표기가 일반화되어 있으면 우리말로 표기한다.

예: 가트(←GATT): 코메콘(←COMECON): 아세안(←ASEAN)

(4) 번역어가 더 빈번히 사용되는 경우에는 번역어를 선택한다.

예: 품질관리(←QC): 판매시점정보관리(←POS)

(5) 한자어와 우리말이 병용되는 경우에는 더 빈번히 사용되거나 일반적으로 통용되는 우리말을 선택한다.

예: 돼지고기(←돈육): 닭고기(←계육)

4.6 상품명의 처리

(1) 적절한 일반명이 있는 경우에는 시소러스에 수록된 일반명을 색인어로 선택한다.

(2) 흔히 탐색어로 사용될 수 있는 유명한 상품명은 자유키워드로 색인한다.

4.7 용어의 표기

용어의 표기 문제는 한글의 표기, 외래어의 표기, 외국어(영어)의 표기로 구분되며, 한글의 표기에는 특히 띄어쓰기 문제가 포함된다.

한글의 표기는 국가가 고시한 표준적인 '한글 맞춤법'에 따른다. 띄어쓰기는 원칙적으로 품사가 다른 각 단어는 띄어쓰도록 되어 있으나 이 원칙을 따른 신문기사

는 거의 없다. 색인에 있어서의 띄어쓰기 문제는 복합명사의 경우 특히 심각해지는데, 보통 빈번하게 사용되는 복합명사는 관용적으로 붙여쓰는 경향이 있다. 따라서 본 연구에서는 색인어로 선정된 복합명사는 모두 붙이도록 하였다. 이렇게 함으로써 다양한 띄어쓰기로 입력된 탐색어들을 하나의 표준적인 형태로 쉽게 변환할 수 있다.

외래어를 우리말로 표기하는 경우 국가가 고시한 '외래어 표기법'에 따라 표기한다. 그러나 일반적으로 신문에서 표기되는 형태를 탐색자가 기억하게 되므로 탐색시 관용적인 형태를 사용할 것이라는 가정을 할 수 있다. 따라서 드물기는 하지만 신문표기가 관용화되어 있는 경우에는 신문표기를 선택하고 표준표기로부터는 참조를 내주도록 한다.

동음이의어는 괄호안에 한정어를 표기하여 식별한다. 한정어로는 한자가 있는 용어와 없는 용어 모두 탐색어 입력의 편의를 위하여 한자를 사용하지 않고 상위개념어나 분야명을 사용하도록 한다.

예: 감자(기업); 감자(농산물)
눈(기상); 눈(신체조직)

5. 경제신문 시소러스의 구축

5.1 시소러스 구축 방향

신문기사 시소러스는 신문기사의 특성과 신문기사 데이타베이스 이용자의 특성을 고려하여 개발하는 것이 바람직하다. 본 연구에서는 다음과 같은 몇가지 사항을 고려하여 시소러스에 수록할 용어를 선정하였다.

(1) 데이타베이스의 이용자가 탐색전문가가 아니고 최종이용자라는 점과 신문기사의 포괄성 및 신문이란 매체의 대중성을 감안하여 전문적이고 학술적인 용어보다는 일반적으로 통용되며 신문에 자주 등장하는 용어를 디스크립터로 선정한다. 또한 관용어, 통칭, 약어 등 다양한 탐색어를 사용할 것을 고려하여 시소러스내에

충분한 참조를 만들어 주도록 한다.

(2) 신문기사에는 유행어나 신조어를 비롯한 일시적인 용어가 많이 나타나므로 이러한 단명어를 디스크립터로 선정해서는 안되며, 오랜 기간 신문기사에 지속적으로 나타날 용어를 디스크립터로 선정한다. 탐색어가 될만한 일시적인 용어는 색인자가 자유키워드로 선택할 수 있다.

(3) 이러한 자유키워드와 시소러스 갱신전에 추가되는 새로운 디스크립터는 별개의 화일에 수록하여 시소러스와 함께 탐색시 사용되도록 한다.

(4) 기관/단체명, 인명, 기업명, 지명 등의 고유명사는 전거화일에 수록하여 색인어 선정시 참조하도록 하고 새로 발생하는 용어는 전거화일에 추가한다. 그러나 제한된 수의 기관/단체명과 지명은 시소러스에 수록하는 것이 편리하다.

(5) 본 연구에서 구축하고자 하는 시소러스는 경제신문을 위한 것이므로 경제관련 분야를 다른 분야에 비해 상세히 개발하도록 한다. 즉 경제와 산업 분야는 충분한 수의 소분야로 나누고 각 소분야에서 심층적인 색인이 가능하도록 색인어의 특정성 수준을 높인다.

5. 2 용어의 수집과 분류

용어의 수집을 위하여 기본적으로는 KETEL이 1991년 1년간 한국경제신문 기사의 색인어로 선택한 2만3천여개의 어휘를 사용하였고, 관련용어의 추가 수집을 위하여는 국내 다른 경제신문들과 종합일간지를 참조하였다. 특히 경제 및 산업 분야의 용어 수집을 위해 중앙경제신문, 서울경제신문, 매일경제신문, 내외경제신문의 기사를 분석하였다.

용어의 주제 분류에는 경제관련 각 전문 분야의 사전과 시사용어사전, 시소러스, 분류표 등을 참고하였다. 특히 언론연구원의 전국언론사 기사자료분류표를 용어분류의 전체적인 틀로 사용하였다. 분류과정에서 의미가 모호한 용어가 많이 발견되었으며 KETEL의 신문기사 데이타베이스를 온라인으로 탐색하여 문제를 해결하였다.

본 연구에서 전체 용어의 대분야 분류는 「전국언론사 기사자료 표준분류표」에 따랐으나 소분야 분류는 관련용어의 수와 주제간의 유사성을 고려하여 조정하였다.

특히 산업분야는 대한통계협회가 발간한 「한국표준산업분류」[8]의 소분야로 나누어 용어들을 분류한 후 표준분류표의 산업을 구성하는 소분야 맞추어 조정하였다. 결과적으로 전체 용어는 10개의 대분야와 46개의 소분야로 분류되었다. 이 가운데 사건사고는 전체안에서의 비중을 고려하여 사회에 포함시키는 것이 나을 것으로 판단되었으나 이용자의 편의를 위하여 표준분류표에 대응하도록 하였다.

5.3 용어의 상호관계 결정

본 연구에서 설정한 용어간의 관계는 동등관계, 계층관계, 연관관계로서 ISO 2788을 기준으로 하여 용어관계를 결정하였다. 동의어는 서로 대체할 수 있을 만큼 의미가 같은 용어들을 말하며, 유사동의어는 일반적인 의미는 다르지만 색인시 동의어로 취급하는 용어들을 말한다.

5.3.1 동등관계

동등관계는 동의어와 유사동의어의 경우 설정하며 USE/UF(Used For)로 연결한다.

(1) 다음과 같은 관계에 있는 용어들은 동의어로 본다.

a. 약어/완전어 (예: GATT/가트/관세무역일반협정)

b. 외래어/번역어 (예: 팀스피리트훈련/한미합동군사훈련)

c. 다른 철자 (예: 로컬가격/로칼가격)

d. 복합어/분해된 형태 (예: 중소기업정보화/중소기업+정보화)

e. 표준명/속칭 (예: 무선호출기/삐삐)

f. 통칭/학술명

g. 일반명/상품명

(2) 다음과 같은 경우에는 유사동의어로 본다.

a. 의미가 중복되는 용어 (예: 금융자율화/금융자유화)

b. 반의어 (예: 경제안정/경제불안; 흡연/금연)

(3) 특정한 개념의 용어를 상위개념어로 색인하도록 두 용어간의 동등관계를 설

8) 대한통계협회, 「한국표준산업분류」 1991.

정한다. 이렇게 함으로써 디스크립터가 되기에는 너무 특정한 개념이라도 무시하지 않고 적절히 색인하도록 한다.

예: 대외자산 UF 대외자산잔고
권력구조 UF 권력구조개편
지역개발 UF 지역균형발전계획

5.3.2 계층관계

계층관계는 종속관계, 계층적 전체-부분관계, 사례관계를 포함하며 관계표시로는 BT/NT(Broader Terms/Narrower Terms)를 사용한다. ISO 2788은 동일한 기본범주(사물, 행위, 행위자, 속성 등)에 속하는 용어들간에만 계층관계를 설정하도록 하고 있다. 즉, 다른 범주에 속하는 용어들간에는 계층관계가 성립되지 않는다는 것인데, 실제로 서로 관련된 용어들을 모아 용어계층을 형성하다 보면 다른 범주의 용어를 한 계층에 포함시키기가 쉽다. 다음의 예에서처럼 '주식'과 관련된 용어들을 모아 계층을 형성할 때 '국민주, 우리사주'들은 주식의 유형으로서 '주식'과 같은 기본범주의 용어들이지만 '주식분할, 주식공개' 등은 다른 기본범주(행위)에 속하는 것들이므로 이 계층에 속해서는 안된다는 것이다.

예: 주식
NT 국민주
공모주
보통주
상장주
우리사주
주식공개*
주식분할*

정확한 계층관계의 설정은 과학기술 분야에서는 비교적 용이하나 인문사회 분야에서는 쉬운 일이 아니다. 실제로 계층관계가 잘 설정되어 있는 INSPEC 시소러스

에서도[9] 다음과 같이 ISO 2788 원칙에 어긋나는 계층을 발견할 수 있다. 이 계층 속에 있는 대부분의 디스크립터는 최상위어인 'fusion reactors'(핵융합반응로)와는 기본범주가 다른 것이거나 그 유형을 나타내는 용어가 아님을 발견할 수 있다.

예: fusion reactors
.fusion reactor ignition
.fusion reactor instrumentation
.fusion reactor materials
..fusion reactor fuel
.fusion reactor operation
.fusion reactor safety
.fusion reactor theory and design
.hybrid reactors

이러한 경우는 인문사회 분야에서는 흔히 발견되는데 「日經シソ-ラス」나 「日刊工業シソ-ラス」에서는 위의 원칙을 거의 따르고 있지 않다. 이러한 사례는 OECD가 발간한 경제 및 사회개발 분야의 시소러스인 「Macrothesaurus」에서도 발견된다.[10] 다음의 'energy' 계층에는 에너지정책, 에너지보존 등 에너지 유형 이외에 에너지와 관련된 용어들이 포함되어 있다.

예: energy
.energy economics
..energy policy
.energy conservation
.energy resources
..biomass energy
..electrical power

9) The Institution of Electrical Engineers, *INSPEC Thesaurus* . Surrey, England: The Gresham Press, 1991.
10) OECD, *Macrothesaurus for Information Processing in the Field of Economic and Social Development*, 1991.

...hydroelectric power
..gas fields
..neuclear energy
..petroleum resources
(이하생략)

그러나 실제로 ISO 2788의 원칙을 따르다보면 서로 밀접하게 관련된 용어들이 한 계층에 속하지 못하고 고립어가 되기 때문에 이러한 용어들이 모두 관련어로 연결되지 않는 한 시소러스 이용자에게 전달하는 정보가 적어지는 문제점이 있다.

다음은 계층관계가 설정되는 각 경우에 관한 설명이다.

(1) 종속관계는 속(屬)/종(種); 범주/구성요소; 사물, 행위, 행위자, 속성의 유형에 적용된다.

예: 노동관계법
NT 근로기준법
남녀고용평등법
노동조합법
노동쟁의조정법

종속관계를 설정할 때 문제가 되는 것은 한 계층 속에 하나의 기준에 의해 분류된 구성요소들만이 포함되지 않고 다른 기준에 의한 제2의 구성요소 집단이 형성될 수 있을 때 둘 이상의 구성요소 집단을 한 계층 속에 열거할 수 있느냐 하는 것이다. 다음 예에서 '경제' 아래 국가별로 분류된 경제유형들이 열거되어 있으며, 다시 '선진공업경제'와 '신흥공업경제'란 다른 측면에서 분류한 경제의 유형이 들어 있다.

예: 경제
한국경제

.세계경제

..일본경제

..미국경제

(중략)

.선진공업경제

.신흥공업경제

실제 시소러스 사례를 살펴보면 한국언론연구원의 시소러스의 경우에는 '경제'의 하위개념어로서 '세계경제'와 '한국경제'만을 포함시키고 '거품경제, 지하경제, 대외경제' 등은 관련어로서 처리하고 있다. 이와 같이 처리할 때에는 한 가지 기준만을 사용하여 경제유형을 분류함으로써 최상위어인 '경제'라는 탐색어를 사용했을 때 각국의 경제와 관련된 기사만 검색할 수 있는 장점이 있지만 반대로 국가별 경제유형이 아닌 다른 경제유형에 관련된 기사를 기대하는 탐색자의 정보요구는 충족시킬 수 없을 것이다.

「Macrothesaurus」에서는 다음 예와 같이 투자의 유형이 될 수 있는 용어들을 'investment'라는 최상위어 아래 모두 열거하고 있다. 여기에는 투자대상별로 구분한 투자유형뿐만 아니라 다양한 유형의 투자가 포함되어 있다.

예: investment

.agricultural investment

.direct investment

.industrial investment

.international investment

..foreign investment

.private investment

.public investment

.securities

(이하생략)

본 연구에서는 한 범주에 속하는 구성요소들은 대부분의 경우 하나의 계층에 포함시켰다. 예외적인 경우의 예로서 '투자'의 경우 투자의 대상이 되는 개념을 포함한 용어(설비투자, 건설투자 등)만을 '투자' 계층에 포함시키고 기업의 해외투자를 의미하는 '해외투자'는 별개의 계층의 최상위어가 되도록 하였다. 다른 예로는 '대학'의 하위개념인 '개방대학, 교육대학, 기술대학, 방송대학, 국립대학, 사립대학' 등은 하위개념어로 설정했으나 입시제도에 의해 구별되는 '전기대, 후기대, 분할모집대'는 관련어로 처리하였다.

이와 같이 용어 계층 형성시 하위개념어와 관련어를 구분하기가 어려운 경우가 많다. 특히 '거품경제', '지하경제'와 같이 상위개념인 '경제'의 한 유형이라기 보다는 특정한 측면을 나타내는 용어를 계층에 포함시킬 것인가는 결정하기 어려운 문제이다. 유사한 경우로는 '수표'와 관련된 '위조수표'와 '부도수표', '기자'와 관련된 '사이비기자', '해직기자' 등이 있다. 본 연구에서는 이러한 다소 부정적인 특성을 나타내는 용어들은 계층에 포함시키지 않고 관련어로 처리하였다. 다음의 예는 이러한 경우 생성된 디스크립터 엔트리이다.

예: 기자
BT 언론인
NT 방송기자
신문기자
특파원
RT 사이비기자
해직기자

(2) 계층적 전체-부분관계는 신체의 부위(예: 신체조직 NT 뇌); 지리적 명칭(예: 프랑스 NT 파리); 학문영역(예: 자연과학 NT 물리학); 계층적 사회구조(예: 군대 NT 사단) 등에 적용된다.

(3) 일반적으로 사례를 나타내는 식별어(고유명사)는 별개의 화일에 유지하며, 식별어를 시소러스에 포함시키는 경우 식별어와 범주를 나타내는 광의어는 사례관계가 된다. (예: 바다 NT 지중해, 대서양, 태평양 등)

5.3.3 연관관계

연관관계는 상호 계층적인 관계는 아니지만 어떤 관련성이 있는 용어들의 관계를 말하며 관계표시기호로는 RT(Related Terms)를 사용한다. 연관관계는 동일한 범주(속)에 속하는 형제용어들간에는 설정할 필요가 없다. 앞의 종속관계에 대한 예에서 '근로기준법', '노동조합법' 등은 '노동관련법'이라는 같은 범주에 속하는 형제용어들이기 때문에 연관관계를 설정하지 않는다. ISO 2788은 다른 범주에 속하는 용어들로서 연관관계가 성립하는 전형적인 경우들을 열거하고 있는데 그 가운데 신문기사 색인어간에 흔히 발견되는 경우는 다음과 같다.

a. 학문영역/연구대상
   예: 정보학 RT 정보시스템, 심리학 RT 청소년심리
b. 과정(작업)/행위자(도구)
   예: 핵실험 RT 핵무기
c. 직업/직업종사자
   예: 그래픽디자인 RT 그래픽디자이너
d. 행위/행위의 결과
   예: 폭력 RT 상해, 출판 RT 도서
e. 행위/행위의 대상
   예: 독서 RT 도서, 교육 RT 학생
f. 사물, 과정, 행위, 행위자 등과 이에 관련된 성질
   예: 장애자 RT 심신장애
   살해범 RT 살인
   방위산업체 RT 무기수출

실제로 연관관계의 설정은 시소러스 작성자의 관점에 따라 좌우될 수 있는 임의적인 요소를 갖고 있다. 또한 연관관계의 과다설정은 시소러스의 부피를 증대시키고 검색효율의 저하를 초래할 수가 있다. ISO 2788에서는 용어들간에 개념적인 관계가 있어서 색인과 검색시 상호 대체용어가 될 수 있는 경우 이들 용어간에 연관관계를 설정하도록 권하고 있다. 즉 하나의 용어가 색인어로 사용될 때 다른 용어

가 강하게 연상된다면 이 두 용어간에 연관관계를 설정하라는 것이다. 실제로 다음의 예에서와 같이 하나의 용어가 다른 용어를 설명하거나 정의하기 위해 꼭 필요하다면 두 용어는 연관관계를 설정할 만큼 밀접하게 관련있는 경우가 많다.

예: 국토개발 RT 개발부담금
금리 RT 금리인하
분양가 RT 분양가격차등화

그러나 위의 '금리인하'의 경우와 같이 한 용어가 다른 용어 속에 내포되는 경우에는 대개 시소러스 편성시 두 용어들이 인접해서 나타나게 되므로 연관관계 표시를 해주지 않아도 관련어를 찾아내는 데 별 어려움이 없다. 따라서 시소러스의 부피를 줄여야 할 때는 이러한 용어들간에는 연관관계를 설정하지 않는 것이 좋을 것이다.

본 연구에서는 하나의 디스크립터가 다른 디스크립터의 앞부분과 완전히 일치하는 경우 두 용어간에는 연관관계를 설정하지 않았다. 즉 '여행'은 '여행대상국, 여행경비, 여행정보' 등의 연관어로 표시하지 않았다. 그러나 '우주선'과 '우주개발, 우주인' 등은 완전한 내포관계가 아니므로 상호 관련어로 표시하였다.

6. 신문기사 시소러스의 활용

6.1 색인작업에서의 활용

실제로 구축된 시소러스는 색인작업에 다음과 같이 사용한다.

(1) 색인대상이 되는 개념을 표현한 용어(즉, 신문기사 속에 출현한 용어)가 자모순 시소러스에 디스크립터로 수록되어 있는 경우에는 이 디스크립터를 색인어로 선택한다. 분야별 계층 시소러스에서는 해당 용어가 속할 주제분야(예: 금융일반)를 먼저 결정하고 다시 그 하위 주제범주를 나타내는 최상위 디스크립터(예: 금융정책)를 찾아서 해당 용어가 수록되어 있는지 확인하여 있으면 색인어로 선택한다. 어떤 디스크립터는 둘 이상의 주제분야에 속할 수 있으므로 해당되는 주제분야를

찾아가도록 한다.

(2) 위의 용어가 비우선어(도입어)인 경우 USE 참조에 의해 디스크립터에 연결되어 있으므로 참조된 디스크립터를 색인어로 선택한다.

(3) 색인 개념을 표현할 수 있는 적절한 용어가 발견되지 않는 경우 분야별 계층 시소러스를 참조하여 상위개념어를 색인어로 선택한다. (예: '고용관리시스템'은 시소러스에 나타나 있지 않으므로 상위개념어인 '고용관리'를 선택한다.)

(4) 특정한 개념의 용어를 디스크립터로 선택하는 대신 상위개념어를 사용하도록 참조를 내준 경우에는 참조된 상위개념어를 색인어로 선택한다.

(5) 특정한 개념의 용어가 두 개 이상의 다른 디스크립터로 분해되는 경우에는 분해하여 색인한다. (예: '중소기업성장률'은 '중소기업'과 '성장률'의 두개의 디스크립터로 색인한다.)

(6) 복합명사의 경우 시소러스를 참고하되 시소러스에 참고할 만한 적절한 엔트리가 나와 있지 않은 경우에는 범위주기에 기술된 지시에 따른다.(예: 경제분야의 디스크립터인 '대외경제협력' 아래 범위주기를 보면「수록되지 않은 두 나라간 경제협력은 '한소경제협력'과 같이 색인하시오」라는 지시가 나와 있다.)

(7) 인명, 기업명 등의 고유명사는 시소러스에 수록되어 있지 않다. 그러나 이러한 식별어들은 일시적이거나 한시적이 아니고 지속적으로 사용되므로 전거화일에 별도로 수록하여 색인어 선정시 참고하도록 한다.

(8) 일시적이거나 한시적인 성격의 위원회, 전시회, 협상, 회의 등의 명칭과 탐색어로 사용될 만한 상품명은 자유키워드로 색인하되 탐색시 참고할 수 있도록 별개의 화일에 수록하여 두는 것이 바람직하다. 자유키워드로 색인하라는 지시는 관련된 디스크립터 아래 범위주기에 명기하였다.

(9) 시소러스에 수록되어 있지 않은 새로운 주제어는 색인자가 판단하여 디스크립터로 사용하고 시소러스 갱신시까지 별개의 화일에 모아두었다가 시소러스에 추가하도록 한다.

6. 2 데이타베이스 탐색시의 활용

신문기사 시소러스를 각 탐색자에게 책자 형태로 제공하여 탐색시 이용하도록 하

기는 힘들다. 대신 시소러스 화일을 구축하여 온라인으로 탐색하도록 함으로써 적절한 탐색어를 선택하는 데 도움이 되도록 하여야 한다. 탐색자에게 제공되는 정보는 입력된 탐색어 전후에 오는 디스크립터들의 리스트와 각 디스크립터 아래 색인된 기사의 수, 각 디스크립터와 어의적 관계를 갖는 다른 용어들이다. 탐색자는 온라인 시소러스를 이용하여 적절한 탐색어를 선정할 수 있을 뿐만 아니라 탐색결과가 만족스럽지 않을 경우 계층관계나 연관관계에 있는 다른 탐색어를 추가로 사용하여 탐색확장을 꾀할 수 있다.

온라인 시소러스의 제공과 함께 검토되어야 할 것은 동의어 사전화일을 구축하여 표준적인 탐색어의 자동선택이 가능하도록 하는 일이다. 시소러스로부터 동등관계에 있는 용어들, 즉 USE/UF에 의해 상호참조되어 있는 용어들을 기초로 하여 별개의 화일을 구축하고 탐색자가 시소러스에 수록되어 있는 디스크립터와 동등관계에 있는 비우선어를 탐색어로 입력하였을 경우 이 화일을 이용하여 해당되는 디스크립터로 자동변환시키는 것이다. 이렇게 함으로써 탐색자는 온라인 시소러스를 탐색하는 번거러움이 없이도 적절한 탐색어를 입력한 결과를 가져올 수 있을 것이다.

7. 결 론

본 연구에서는 전문을 대상으로 하는 신문기사 색인시스템의 성능을 높이는 방안을 모색한 결과 현재 사용중인 통제되지 않은 수작업 키워드색인 방식과 자동색인 방식이 모두 비효과적이라는 결론을 얻었다. 이에 대한 우선적인 대안은 시소러스를 이용한 통제언어 색인으로서 이 색인 방식은 색인어 선정원칙을 제공함으로써 색인의 일관성을 유지할 수 있도록 하며 색인어 선정시 발생하는 색인자의 고충을 덜어준다. 또한 표준화된 색인어를 부여하게 되므로 동일한 주제의 기사들을 한곳에 모아주며 탐색시 적절한 탐색어의 사용을 통해 높은 검색효율을 기대할 수 있게 된다.

따라서 본 연구에서는 신문기사 색인을 위한 우리말 키워드의 선정원칙을 수립하고, 실제로 한국경제신문사가 제공한 2만여개의 어휘와 새로 수집한 상당수의 어

휘를 대상으로 하여 시소러스를 개발하는 과정에서 제기되었던 여러 문제들을 검토하였다. 이와 같이 시소러스 개발 사례를 통해 밝혀진 구체적인 절차와 이에 관련된 문제점들은 앞으로 다른 시소러스 개발에 참고할 수 있을 것이다.

참고문헌

대한통계협회, 「한국표준산업분류」 1991.

한국언론연구원, 「신문기사 종합시소러스: Kinds 신문기사정보 검색용어집」 1993.

日刊工業新聞社, 「日刊工業シソ-ラス」 1990.

日本經濟新聞社, 「日經シソ-ラス」 1990.

中日新聞本社, 「ニユ-ス・シソ-ラス」 1990.

Aitchson, Jean and Alan Gilchrist, *Thesaurus Construction: A Practical Manual*, 2nd ed. London: Aslib, 1987.

Houston, J.E., ed., *Thesaurus of ERIC Descriptors*, 10th ed. Phoenix: Oryx Press, 1984.

The Institution of Electrical Engineers, *INSPEC Thesaurus*. Surrey, England: Gresham Press, 1991.

International Organization for Standardization, *ISO 2788: Guidelines for the Establishment and Development of Monolingual Thesauri*, 2nd ed. Geneva: ISO, 1986.

OECD, *Macrothesaurus for Information in the Field of Economic and Social Development*, 1991.

Office of Naval Research, *Thesaurus of Engineering and Scientific Terms*, 1967.

UNISIST Guidelines for the Establishment and Development of Monolingual Thesauri, 2nd ed. Paris: Unesco, 1981.Surrey, England: The Gresham Press, 1991.

ABSTRACT

Newspaper Thesaurus Construction in Theory and Practice

Young-Mee Chung

Effective indexing systems are required to enhance the performance of full-text retrieval systems. The result of the analysis of index terms selected by human indexers without a newspaper thesaurus indicates that controlled indexing language is necessary for effective and consistent indexing of newspaper articles. In this paper, basic principles are established for keyword selection from Korean newspapers and significant problems identified in the process of developing a newpaper thesaurus are discussed in depth.

* Professor, Dept. of Library and Information Science, Yonsei University.