용어정의와 관계추출을 통한 시소러스 확장에 관한 연구*

A Study on Thesaurus Expansion through Definitions of Terms and Extraction of Relationships

김 지 훈(Ji-Hun Kim)**
김 태 수(Tae-Soo Kim)***

목 차

초 록

정보검색과정에서 용어의 일관성을 유지하기 위해, 시소러스에서 용어의 의미를 명확하게 제시하는 것이 필요하다. 이에 대부분 시소러스는 기본관계나 범위주기를 이용하여 용어의 의미를 제시하여 왔으나, 최근 내용과 형식에 있어서 표준화된 정의를 포함하는 시소러스가 제안되어 왔다. 이 연구는 표준화된 정의를 작성함과 동시에 그러한 과정에서 사용한 정의모델의 내용에서 관계를 추출하고, 이들 표준정의와 추출된 관계를 기존의 시소러스에 통합하거나 대체하여 확장된 시소러스를 구축해 봄으로써, 시소러스가 더욱 발전할 수 있는 가능성을 고찰하였다.

ABSTRACT

To maintain consistency of terms in information retrieval process, it is necessary to present the meaning of terms definitely in thesaurus. Therefore, most of thesauri has presented meaning of terms through basic relationships or scope notes. But, thesaurus including standardized definitions in contents and form has been proposed lately. This study was performed to make standardized definitions and extract relationships in contents of defining models. Also, expanded thesaurus was constructed being integrated and replaced standardized definitions and extracted relationships into the existing thesaurus. As the result, this study has shown a possibility for further development of thesaurus.

키워드: 시소러스 확장, 전문용어 시소러스, 표준정의
Thesaurus Expansion, Terminological Thesaurus, Standardized Definitions

---

* 이 연구는 연세대학교 대학원 박사학위논문의 일부를 요약한 것임.
** 계명문화대학 사회복지상담과 부교수(jhkim@km-c.ac.kr)
*** 연세대학교 문헌정보학과 교수(btree@yonsei.ac.kr)
논문접수일자 2006년 2월 15일
게재확정일자 2006년 2월 16일

1. 서 론

정보검색은 기본적으로 색인과정과 탐색과정에서 선정한 용어들 간의 일치여부에 기반을 두고 있다. 그러나 색인자와 탐색자가 동일한 개념을 나타내는데 종종 서로 다른 용어를 사용하기 때문에 검색시 서로 일치하지 않아 검색효율을 떨어뜨리는 원인이 되어왔다. 이러한 문제를 해결하기 위해, 정보검색에서는 단순히 용어의 존재유무에 근거하는 용어기반검색이 아니라 용어의 개념과 그 관계에 근거한 개념기반검색을 지향하여 왔다.

개념기반검색은 색인과정에서 문헌을 대표하기 위해 선정한 용어와 탐색과정에서 정보요구를 나타내기 위해 선정한 용어가 서로 일치하지 않더라도 동일한 개념이면 서로 일치되도록 하는 것으로, 정보검색에서 이를 위해 사용해 온 주요한 보조도구가 시소러스다.

시소러스의 역할은 정보검색과정에서 주제를 나타내는 용어 및 개념간의 관계를 제시함으로써 용어를 선정하는데 도움을 주어 정보검색의 품질을 향상시키는 것이다. 이를 제대로 수행하기 위해 시소러스는 기본적으로 특정 주제분야에서 사용되는 용어의 의미뿐만 아니라 그 주제 분야의 구조를 이해하는데 도움이 되도록 용어간의 관계를 명확하게 제시하여야 한다.

오늘날 대부분의 시소러스는 ISO 2788, ISO 5067, ANSI/NISO Z39.19와 같은 표준이나 시소러스 구축에 관한 실제적인 지침에서 제시하고 있는 관계구조에 근거하여 용어의 의미와 용어간의 관계를 제공하고 있다. 그러나 각 주제마다 용어의 수가 증가하여 시소러스가 양적으로 팽창해짐에 따라, 기존의 단순한 관계구조를 통해서는 용어의 의미를 명확하게 파악하기 어려워지게 되었을 뿐만 아니라 용어간의 정밀하고 다양한 관계를 제시하기 어려워지게 되었다.

이에 시소러스 연구자들은 시소러스에서 용어간의 관계를 확장하는 것뿐만 아니라 용어의 의미를 제시하는 문제에 많은 관심을 가지게 되었다. 그 결과, 관계구조의 확장에 대해서는 여러 방안들이 다양하게 제시되어 왔지만, 용어의 의미를 제시하는 부분은 동등, 계층 및 연관관계 등의 기본관계나 범위주기와 같은 부가관계를 통해 간접적으로 제시하는 것 이외에 용어가 지니고 있는 개념의 의미를 직접 제시하는 방법이 제시되지 않아 왔다.

그러나 용어의 의미를 제시하기 위해서 사용해 온 관계구조는 용어의 개념이 지닌 본질적인 속성을 이해하거나 범위를 한정하기 어려운 면이 있다. 또한 최근 시소러스에 수록되는 용어의 수가 증가함에 따라 개념간의 충돌이나 모호한 의미가 존재하는 등의 문제로 인해 용어의 의미를 정확하게 확인하는 것이 점차 어려워지고 있다. 이런 이유로 인해, 시소러스에서 관계구조를 더욱 자세히 설정하는 방법뿐만 아니라 용어의 의미를 더욱 명확히 제시할 필요성이 지속적으로 제기되어 왔다.

사실 정보검색과정에서 용어로 표현되고 있는 개념의 의미가 명확해야 용어의 일관성이 유지되며 용어간의 관계를 분명히 할 수 있다. 용어의 의미가 명확하게 파악되지 않으면, 정보검색과정에서 정확성을 떨어뜨리거나 비일관성을 야기하여 정보검색의 품질을 떨어뜨릴 수 있다. 이것은 용어의 의미를 통해 시소러스에 수록할 용어의 선정기준을 확립하고, 선정된 용어의 범위를 한정하고, 아울러 관계구조

를 보다 완벽하게 표현할 수 있음을 의미한다.

이처럼 용어의 의미가 시소러스에서 중요하게 인식됨에 따라, 일부 시소러스는 디스크립터로 선정된 용어의 의미를 제공하기 위해 그 용어의 정의를 함께 수록하고 있지만, 그 정의의 내용이나 형식이 만족스럽지 않은 것으로 제기되어왔다. 물론, 시소러스에서 내용이나 형식면에서 표준화한 정의를 체계적으로 포함하려는 생각은 시소러스를 개발한 이래 여러 범 표면화되어 왔다. 그러나 그러한 시도는 이론적인 면에서 확고한 원칙과 절차가 없었을 뿐만 아니라 정의 그 자체가 시소러스에서 어느 정도 유용한지에 대한 명확한 해답을 도출하지 못해 실제로 적용되지는 않았다.

이에 본 연구에서는 전문용어(terminology) 분야에서 정의를 작성하는데 이용되고 있는 정의모델과 정의규칙에 의거하여 내용과 형식면에서 일관성 있는 용어의 표준정의를 작성하고, 또한 그러한 표준정의를 작성하는 과정에서 사용한 정의모델의 내용에서 관계를 추출하였다. 나아가 새로 작성한 표준화된 정의와 추출된 관계를 기존 시소러스에 부가하거나 대체하여 더욱 확장된 시소러스를 프로토타입으로 구축하여 봄으로써 시소러스의 발전 가능성을 고찰하였다.

2. 시소러스에서 용어의 정의와 활용

2.1 용어의 정의와 시소러스에서의 사용

2.1.1 용어, 개념 및 정의의 이해

'용어'는 기본적으로 특정 주제에서 사용되는 하나의 개념을 유일하게 나타내는 주요한 단위로서, 단일어 또는 복합어로 된 어휘를 일컫는다. '개념'은 동일한 특성을 가진 대상을 추상화하여 일반화한 관념으로, 이를 기호로 나타낸 것이 용어다. 이를 통해, 개념은 용어의 의미와 동일하며 하나의 용어로만 표현되어야 함을 알 수 있다. 또한 개념은 명확하게 설명될 수 있어야 하는데, '정의'는 그러한 설명을 기호 즉 용어로 표현한 것이다(Sager 1990). 요컨대, 용어는 특정 주제에서 정의를 가진 하나의 개념이라 할 수 있다.

그러나 일반적으로 용어는 문맥에 따라 여러 가지 개념으로 나타나기 때문에 본질적이고 분명한 의미를 파악하기 어려우나, 정보검색에서는 시소러스를 이용하여 이를 일부 해결하여 왔다. 즉, 시소러스는 용어와 개념 간에 일대일(one-to-one)관계를 유지해야 하는 용어 데이터베이스라 할 수 있다. 그러나 시소러스가 "상위 및 하위개념 사이의 관계를 명백하게 하기 위하여 공식적으로 조직·제어된 색인어의 어휘집"이고 "색인어는 명사나 명사구의 형태로 나타낸 개념의 표현"이라 정의된 것처럼(International Organization for Standardization 1986), 시소러스는 개념을 주요 요소로 간주하지만, 개념을 나타내는 용어가 자연언어에서 채택되는 본질적인 문제로 인해 용어와 개념간의 관계를 분명하게 구별하지 않고 있다. 따라서 개념은 명확히 정의되어야 할 뿐만 아니라 용어와의 관계도 분명히 구별되어야 할 필요가 있다.

용어의 의미 즉 개념을 정의하기 위해서는 개념을 추상화하는 것과 함께 개념의 특성을 파악해야 한다. 개념의 추상은 개념의 대상물에서 서로 다른 특성을 제거하고 같은 특성만

종합하여 유-종(genus-species)관계를 만드는 과정이다. 따라서 개념간의 계층은 그 추상수준에 따라 일반적으로 인접한 상위개념에 따라 정렬되며, 관계유형에 따라 하위개념을 가지는 피라미드 형식으로 만들어진다. 이로써 하나의 개념은 인접한 상위개념의 특성과 동일수준의 다른 개념과 구별되는 그 개념만이 가지는 특성으로 정의될 수 있다.

한편, 개념의 특성은 어떤 개념을 구별하기 위한 개념의 최소 요소로 간주되어 개념을 기술하고 분류하고 정의하는데 사용하지만, 제시되는 모든 특성 중 어떤 특성이 적합한지 판단하기는 쉽지 않다. 그러나 개념의 특성 중에서 가장 적합한 특성만을 포함한 것이 정의가 되기 때문에 개념의 특성을 파악하는 것은 무엇보다 중요하다. 일반적으로 개념의 특성은 내포(intension)와 외연(extension)으로 구분하고 있다(Felber 1984).

내포는 개념 간에 구별을 위한 본질적인 속성이며, 외연은 개념이 속하는 동일한 추상수준의 모든 개념 또는 모든 개별대상의 집합을 의미한다. 이러한 구분에 따라 개념을 정의하는 방법은 내포적 정의와 외연적 정의로 구분되어, 전자는 개념의 내포 즉 구별특성들을 기술하는 것이고, 후자는 그 개념과 동일한 추상수준의 개념을 열거하는 것이다.

일반적으로 언어사전의 정의는 단순히 단어의 의미를 기술하는 반면, 전문용어의 정의는 다른 개념과 구별되는 모든 특성에 근거하여 그 의미를 기술하기 때문에, 내포적 정의는 전문용어를 정의하는데 가장 이상적인 정의로 인식되고 있다(Sager 1990). 내포적 정의는 분석적 정의로도 알려져 있는데, 그 구조는 아리스토텔레스의 '피정의항 = 정의항(인접류 + 구별특성)'이라는 정의원칙에 기반을 두고 있다(Dahlberg 1981). 즉 용어를 분석적으로 정의하는 것은 그 개념이 속한 유와 그 유에 속한 다른 개념과 구별되는 특성인 종차(differentia)를 나타내는 것이다(Sager and Ndi-Kimbi 1995).

2.1.2 시소러스에서 용어의 정의

시소러스에서 용어와 개념의 관계는 일대일 관계가 이상적이지만, 개념은 자연언어를 이용하여 나타내기 때문에 용어와 일대일관계가 유지되지 않는 문제가 발생한다. 그럼에도 불구하고 지금까지 이것이 시소러스에서 그다지 문제시되지 않은 것은 시소러스에서 의미를 다루려고 하는 의도가 없었기 때문이기 보다는, 시소러스의 용어가 다른 용어와의 관계 그 자체로 정의되거나, 시소러스가 전문용어를 다루고 있을 뿐 아니라 그 주제 분야의 전문가들만 이용하기 때문에 의미를 동일하게 인식하고 있을 것이라 생각한 것으로 보인다(Soergel 1974; International Organization for Standardization 1986; Buchan 1989).

그러나 시소러스에서 용어의 의미를 명백히 하기 위해 용어의 정의를 제시할 필요가 있는데, 대부분의 시소러스는 용어의 의미를 실제적이고 완전하게 제시하지 않고 간접적으로 제시하여왔다. 간접적인 방법으로는 주로 시소러스 표준이나 지침에서 정해진 기본관계나 부가관계로서 범위주기를 이용하여 왔으며, 최근 일부 시소러스에서는 범위주기와의 차별성을 강조하여 정의를 부가하고 있다.

시소러스에서 기본관계는 개념이 동등하다면 용어 간에 동등관계를 가지고, 개념이 다른

개념보다 상위개념 또는 하위개념이라면 용어 간에 계층관계를 가지며, 개념이 의미적으로 관련 있으면 용어 간에 연관관계를 가지는 것으로 인식한다(Svenonius 1990). 이러한 기본 관계는 '개념간의 관계'와 '용어와 개념간의 관계'로 구분하여 다루고 있는데, 그 내용은 다음과 같다(Soergel 1985).

개념간의 관계는 여러 개념 중에서 하나 이상의 공통적인 특성으로 만드는데, 형식적(formal) 관계와 실체적(material) 관계로 구분한다(Dahlberg 1989). 이 중 실체적 관계는 계층관계, 분할관계, 대립관계, 기능관계 등 네 가지로 구분하고 있는데, 이를 통해 용어의 의미를 나타내는 정의를 다음과 같이 설명할 수 있다. 계층관계를 이용한 정의는 개념을 나타내는 상위어와 하위어가 논리적인 배열로 제시된 것으로, 유-종관계로 다루어져 분석적 정의형태로 나타난다. 분할관계를 이용한 정의는 개념을 대상의 구성요소로서 정의하는 것으로, 정의에서 언급된 전체개념은 상위개념과 동일하므로 계층관계의 일부분에 해당된다. 대립관계를 이용한 정의는 개념이 부정 또는 반대로 정의되는 것이고, 기능관계를 이용한 정의는 전체개념의 요소 및 특성을 포함한다. 이처럼 실체적 관계는 개념이 반대로 정의되는 대립관계 외에 개념에 대해 많은 의미를 포함하고 있음을 볼 수 있으며, 일반적으로 대부분의 일반사전에서는 계층관계와 분할관계를 이용하여 정의하는 반면, 전문사전은 기능관계를 이용하여 정의하고 있다.

용어와 개념간의 관계는 등가관계와 연관관계를 들 수 있다. 등가관계는 동의어나 유사동의어가 해당되는데, 시소러스에서 동일한 개념이라고 인정되는 경우에는 우선어와 비우선어 간의 관계로 파악하여 USE와 UF로 나타내고 있다. 그러나 시소러스에서 동의어를 이용하여 용어의 의미를 설명하는 것이 일반적이지만, 개념이 용어의 동의어로 정의될 수 없다는 Sager(1990)의 지적처럼, 등가관계는 기본적으로 개념과 용어 간에 일대일 관계가 아닐 뿐만 아니라 분석적 정의형태로 개념을 정의할 경우에도 적용될 수 없다. 예컨대 '새'를 '조류'로 정의할 수는 없는 것과 같은 이치다.

또한 연관관계는 개념적으로 밀접하게 관련되어 있으나 등가관계에 포함되지 않는 관계로서, 개념 간에 명시적으로 관련 지워야 할 정도로 의미적으로나 심리적으로 연관이 있는 경우에만 만들도록 권고되고 있는 만큼(International Organization for Standardization 1986; National Information Standards Organization 1994), 명확하게 규정하기가 대단히 어렵다. 그러나 용어가 다른 용어의 정의나 설명에서 필요한 요소로 보여 지면 연관된 것으로 고려하는 것처럼(Aitchison and Gilchrist 1987; Soergel 1974), 연관관계는 용어의 의미설명을 일부분 담당한다고 볼 수도 있다.

이상의 내용으로 보아 시소러스의 기본관계는 용어의 의미를 간접적으로 제시하기는 하지만, 그 관계가 개념간의 관계인지 용어간의 관계인지가 분명하지 않으며(McNaught 1983), 용어의 의미를 관련어들로 파악하는 것이 다소 무리가 있으며(Moores 1985), 특정 주제에서 개념간의 상호관계에 대한 포괄적인 이론이 아직 없다는(Nkwenty-Azeh, 1994) 등의 지적이 있어 왔다. 특히 Sager(1990)는 디스크립터의 의미를 결정하는데 기본관계를 통해 약간의

도움을 받을 수는 있겠지만, 개념을 세 가지 기본관계로만 의미를 나타내는 것은 극도로 단순하다고 지적하였다. 이러한 한계를 극복하기 위해 시소러스에서는 범위주기를 주로 사용하고 있으며, 근래에 와서는 이와 별도로 정의를 점차 포함하고 있다.

일반적으로 범위주기는 디스크립터의 의미를 명백하게 설명하거나 이용법을 제시할 필요가 있을 때 제공된다. 범위주기에 대해 시소러스 표준에서는 "디스크립터의 활용을 제한 또한 확장하거나, 자연언어에서 의미가 중복되는 디스크립터를 구별하거나, 색인자나 탐색자에게 용어를 사용하는데 조언하기 위해 사용하며 나아가 그 의미를 제시해야 하는 것"이고(National Information Standards Organization 1994), 시소러스 개발지침에서는 "색인시 용어를 이용하는 방법에 대한 정보를 제공하는 색인지향주기"로 보고 있다(Aitchison, Gilchrist & Bawden 2000). 그러나 시소러스에서 범위주기는 체계적으로 사용되지 않거나 누락되는 경향이 있으며, 한편으로는 범위주기가 용어에 대한 다양한 정보를 제공하기 때문에 그 기능이 너무 과다하다는 지적도 있어 왔다(Sager, Sommer and McNaught 1982). 따라서 용어의 의미와 사용법의 중요성에 의거하여 범위주기와 구분되는 더욱 완전한 정의를 시소러스에 통합하는데 대한 관심이 증가하였다(Svenonius 1997; Hudon 1996, 1997).

시소러스에 정의를 포함하려는 생각은 시소러스를 처음 만들 때부터 지속적으로 제기되어 왔으나(Soergel 1974; Moores 1985; Buchan 1989; Svenonius 1997), 일부 시소러스를 제외하고는 최근까지 정의를 거의 포함하지 않고 있다. 그 이유는 용어의 정의를 제시하는 것이 그다지 어렵지 않을 것이라는 생각으로 시소러스 개발에 관한 표준이나 지침에서 정의와 범위주기를 구별하여 이용하는 것을 권고하지 않았기 때문으로 보인다.

그러나 "용어의 의미를 위해 시소러스에 범위주기와 정의를 부가하면 시소러스가 더욱 효율적"이고, "시소러스에서 사전정의와 같은 완전한 정의가 제시되지 않았으나, 일부 제시된 정의는 종종 시소러스 관계구조에 의해 전달된 의미를 보충하는데 필요하다"고(Aitchison and Gilchrist 1987; Aitchison, Gilchrist and Bawden 2000) 인정하는 것처럼, 시소러스에서 정의의 필요성과 함께 범위주기와 구별된 정의의 필요성이 제시되어 왔다.

이러한 인식의 변화에도 불구하고, 이들 표준이나 지침에서는 정의를 사전적 의미로 국한하거나 정의가 용어의 의미를 보충적으로 설명하는 역할로서만 제시하였을 뿐 구체적인 지시사항은 없었으며, 또한 정의의 이용을 명확하게 권고하지 않고 있다. 다만 일부 시소러스만이 사전과 통합하거나 범위주기와 구별되는 정의를 부가하는 변화된 모습을 보여주고 있다.

한편, 전문용어분야에서는 시소러스가 그들의 연구내용과 부분적으로 관련 있음을 인식하고 오히려 시소러스에서 용어의 의미를 설명하는 문제에 더욱 관심을 두었다. Sager(1990)는 시소러스에서 용어의 의미를 나타내기 위해 용어 자체, 다른 개념과의 관계 및 정의를 이용할 수 있는데, 그 중에서 정의가 가장 정확하고 유용한 서술적 묘사를 제공한다고 보고 용어간의 의미관계뿐만 아니라 표준화된 정의를 포함하는 전문용어 시소러스의 개발을 제시하였다.

2.2 전문용어 시소러스와 정의작성

2.2.1 전문용어와 시소러스의 관계

전문용어에 대한 연구는 특정 주제 분야의 효율적인 커뮤니케이션을 위해 개념과 용어 간에 나타나는 여러 가지 문제를 체계적으로 연구하고 관련된 각종 정보를 조직하는 것을 목적으로 하고 있다. 이에 전문용어분야에서는 그 연구내용과 방법이 시소러스의 방법론과 서로 유사하다고 인식하고, 전문용어와 시소러스를 통합하여 연구할 것을 권고하여 왔다(Sager 1990). 그 예로써, 전문용어에서 기본단위인 용어와 시소러스에서 기본단위인 디스크립터가 다음과 같은 동일한 특성을 가지고 있다고 보았다.

- 용어와 디스크립터는 하나의 주제에서 하나의 개념을 표현한다.
- 용어와 디스크립터는 자연언어로 표시된 기호다.
- 용어와 디스크립터는 전문분야에서 만들어진 언어형식을 반영한다(Larivière 1989 In Hudon 1998).

또한 〈그림 1〉과 같이 용어은행(termbank)과 시소러스의 구조를 서로 비교하여 제시하고, 올림말(entry)의 기술방법과 정의를 포함하는 전문용어부문이 관계구조를 포함하기 있는 시소러스부문과 큰 차이를 보이고 있지만 서로 부족한 부분을 수용하여 통합한다면 정보검색에서 더욱 유용한 도구가 될 수 있다고도 보았다(Hudon 1998).

그러나 문제는 시소러스에 어떤 정의를 포함하느냐 하는 것인데, 전문용어분야에서 다루어온 '전문용어정의(terminological definition)'가

용어은행 레코드	시소러스 레코드
• *Entry*	• *Descriptor*
• *Subject field*	• *Class*
• Grammatical label	
• Language code	
• Country code	
• Acceptability rating	
• Context	
	• Broader descriptors(BTs)
	• Narrower descriptors(NTs)
	• Associated descriptors(RTs)
• *Definition*	• *Scope note*
• *Note*	
• *Abbreviations*	
• *Synonyms*	• *Synonyms* (UF)
• *Spelling variants*	
• Incorrect forms	
• *Translation equivalents*	• *Linguistic equivalents*
• *Reference / Sources*	• *Sources*

〈그림 1〉 용어은행과 시소러스의 레코드 구조 비교

효과적인 것으로 인식되고 있다. 특히 Sager (1982)는 전문용어정의를 포함한 시소러스를 기존의 시소러스와 구별하여 '전문용어 시소러스'라는 이름으로 개발하는 것을 제안하였다.

2.2.2 전문용어정의

정의는 "주로 언어적 수단으로 표현된 알고 있는 개념의 포괄적이고 서술적인 기술"(International Organization for Standardization 1987)로서, 개념의 영역을 한정하여 동일한 용어를 이용하는 사람들이 동일한 개념을 가지게 하기 위한 것이다. 일반적으로 언어사전의 정의는 모든 대상의 일반적인 경험을 참조하여 단어 그 자체의 모든 의미를 기술하는 반면, 전문용어정의는 전문분야에서 전문적인 경험에 근거하여 실세계에 존재하는 대상의 추상적 표현인 개념을 기술한 것이다(Sager 1990).

언어사전의 정의유형으로 가장 일반적인 것은 다른 단어로의 대치에 의한 정의, 의미론적 정의, 분석적 정의 등 다양하게 있다(이병근 1992). 그 중 분석적 정의는 대상이 가지는 논리적인 유-종관계와 종차를 확인하여 의미를 설명하는 것으로, 전문용어 체계의 기초로 인식되는 "피정의항 = 정의항"과 부합되어 가장 이상적인 전문용어정의의 유형으로 보고 있다(Ndi-Kimbi 1994).

논리적 정의라고도 일컬어지는 분석적 정의는 대상이 되는 개념의 공간상 상대적인 위치를 한정하여 개념이 속한 인접한 유(genus proximus)와 개념의 내포적인 지식인 종차(differentia specifica)를 언급하는 것으로(Dahlberg 1981; Sager and Ndi-Kimbi 1995), ISO 1087-1이나 ISO 704에서도 분석적 정의방식에 기초한 정의형식을 규격화하여 권고하고 있다. 이러한 분석적 정의를 우리말 구조로 보면 '피정의항 = 정의항(종차 + 유개념)'의 형식으로 된다. 예컨대, '바이트(byte)'에 대한 정의가 '8비트로 구성된 문자열'이라면, 정의항에서 '문자열'은 유개념어가 되고 '8비트로 구성된'은 종차가 된다. 여기서 특히 유개념어는 대상이 되는 개념에 최근접한 유개념어가 되어야 한다.

이처럼 분석적 정의형식으로 만들어진 전문용어정의는 특정 주제에서 개념 간에 존재하는 차이를 분명히 하는 특성들에 근거하여 그 개념의 의미를 전달하는 개념의 언어적 기술이다. 이에 Nkwenti-Azeh(1994)는 전문용어정의가 내용이 풍부하고, 그 정의를 분석하여 기본지식구조를 구성할 수 있을 뿐만 아니라 그 정의에서 추출한 용어와 관계구조를 이용하여 전문용어 시소러스를 구축할 수 있다고 보았다.

따라서 시소러스를 구축하는데 전문용어정의에 관심을 가지는 이유는 전문용어정의가 특정 주제에 의존하여 개념의 특성을 설명하고, 용어와 개념간의 관계를 만들어 개념이 속한 체계에서 그 위치를 파악하는 데 사용할 수 있는 많은 관계를 포함하고 있기 때문이라 할 수 있다. 이에 전문용어정의를 작성하기 위한 규칙과 모델이 다음과 같이 연구되고 구체화되어 왔다.

정의작성규칙에 대한 필요성은 많이 논의되었지만(Dahlberg 1981; Ndi-Kimbi 1994; Sager and L'Homme 1994), 전문용어분야에서도 정의에 대한 적절한 내용과 방법에 대한 규칙이 명확하게 제시되지 않았었다. 다만 근래에 와서 새로운 개념을 정의하기 위한 요구가 증가함에 따라 다음의 여러 표준에서 전문분야의

개념에 대한 정의규칙을 정하여 왔다.

- ISO Technical committee 46. Draft Procedures - Annex E: Guidelines for Definitions
- ISO Recommendation R704:1968. Naming Principles
- ISO 704:1987. Principles and Methods of Terminology
- BSI BS 3369:1963. Recommendations for the Selection, Formation and Definitions of Terms

이러한 표준들은 정의하려는 목적과 상황에 따라 정의작성규칙이 다양하게 적용되겠지만, 그 내용은 '형태구문적 조건'과 '어휘의미적 조건'으로 구분되어 제시되었다(Ndi-Kimbi 1994). 이러한 정의작성규칙은 동일한 부류의 개념에 대해 정의패턴을 일치시키고 향후 정의를 작성하기 위한 모델설계를 가능하게 하였다.

한편, 정의를 표준화 할 필요성은 전문용어 분야에서 많이 논의되었다. 정의를 표준화한다는 것은 정의모델을 통해 정의양식을 일치시키는 것으로, 항상 같은 순서로 용어의 본질적인 요소를 파악하여 정의를 쉽게 구성하고 조직할 수 있게 한다는 것이다. 이에 Sager와 L'Homme (1994)는 기존의 대부분의 정의가 데이터베이스 환경에서 효율적으로 이용하는데 적절하지 않다고 보고, 정의를 표준화하기 위해 분석적 정의구조로 정의를 기술하는 양식을 정규화 하여 기존 용어의 정의를 표준화하는 정의모델을 제시하였다.

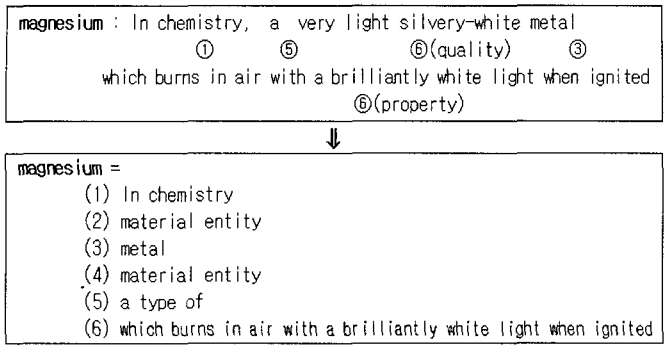
이들의 정의모델은 일관성 있고 완전한 정의를 만들기 위해 일반사전의 항목보다 더욱 엄격히 구분되게 항목을 구분하고 조직하여 기존의 여러 정의를 다시 결합하여 재작성하는 것을 체계화하도록 한 것으로, 데이터베이스 모델을 염두에 두고 〈그림 2〉와 같이 정의에 본질적으로 필요한 7가지 항목이 포함된 정의모델을 제시하였다. 아울러 그들은 제시한 정의모델을 이용하여 〈그림 3〉과 같이 기존의 정의를 각 항목으로 분해하고 그 항목에 근거하여 정의를 재작성할 수 있음을 보여 주었고, 나아가 이것이 데이터베이스에 저장될 때 개별항목에 대한 검색을 가능하게 할 것이라 제안하였다.

2.3 정의에서 의미관계추출

자연언어처리 분야에서 용어간의 의미관계

피정의항 =
(1) 주제분야
(2) 피정의항의 개념범주
(3) 정의항(최근접 유개념어)
(4) 정의항(최근접 유개념어)의 개념범주
(5) 피정의항과 정의항(최근접 유개념어)과의 관계
(6) 피정의항과 관련개념을 구별하는 특성, 즉 종차
(7) 이용자에게 유용할 것으로 보이는 비본질적 특성

〈그림 2〉 Sager와 L'Homme의 정의모델



〈그림 3〉 Sager와 L'Homme의 정의모델 적용사례

를 얻기 위한 방법은 상향식(bottom-up) 방법과 하향식(top-down) 방법으로 구분된다(Condamines and Rebeyrolle 2001). 이 두 가지 방법 중 어느 것도 의미관계를 완전히 추출할 수 없는 것으로 밝혀졌지만, 여러 가지 기법으로 연구되어왔다.

상향식 방법은 데이터에 대한 선행지식 없이 코퍼스내의 데이터에 근거하여 정보를 추출하는 방법으로, 자율방법(unsupervised approach)이라고도 한다. 이 방법은 주로 용어추출모듈과 미리 정의된 용어형태를 가지고 유사한 문맥에 근거하여 용어간의 상관관계(correlations)를 찾는다. 예컨대 구문분석과 통계적 방법을 결합하여 동사의 주어와 목적어를 조사하여 유사한 의미를 가진 단어를 추출하는데, Hindel (1990)은 이 방법으로 boat와 가장 유사한 단어로 ship, plane, bus, jet, vessel, truck, car, helicopter, ferry를 파악하였다. 상향식 방법은 주제지식이 필요하지 않을 뿐만 아니라 단일어간의 관계를 추출하는 효과적이라 평가되었지만, 의미관계가 명시되지 않는 것이 문제로 지적되었다.

하향식 방법은 추출될 데이터에 대해 미리 수작업으로 정한 어휘구문패턴에 근거하여 용어간의 관계를 추출하는 방법으로, 지도방법(supervised approach)이라고도 한다. 이 방법은 주로 자연언어처리에서 주요한 정보원인 사전에서 의미정보를 추출하기 위해 기계가독사전의 정의를 대상으로 많이 사용되었는데, 특정 주제 분야에서는 좋은 성능을 발휘하지만 비용이 많이 든다는 것이 단점으로 지적되었다.

사전정의에서 어휘관계를 추출하기 위해 패턴을 이용한 방법은 Amsler(1980)에 의해 처음 시도되었다. 그는 정의가 정의되는 단어의 상위개념을 나타내는 유개념어와 다른 개념과 구별하는 종차로 구성되어 있다는 가정에 근거하여, 사전에서 유개념어를 수작업으로 추출하여 계층관계로 구성된 시소러스를 만드는 가능성을 보여주었다(Michiels and Noël 1982). 또한 이와 유사하게 사전의 정의를 분석하여 유개념어 뿐만 아니라 구문 및 어휘패턴을 조사하여 의미관계를 추출하는 여러 연구가 있었다 (Chodorow, Byrd, and Heidorn 1985; Markowitz, Ahlswede and Evens 1986).

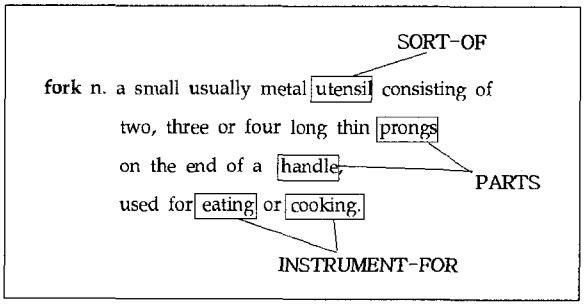
이처럼 의미관계를 추출하기 위해 사전을 이용한 연구는 크게 다음의 두 가지 원리에 근거

한다(Ide and Véronis 1993).

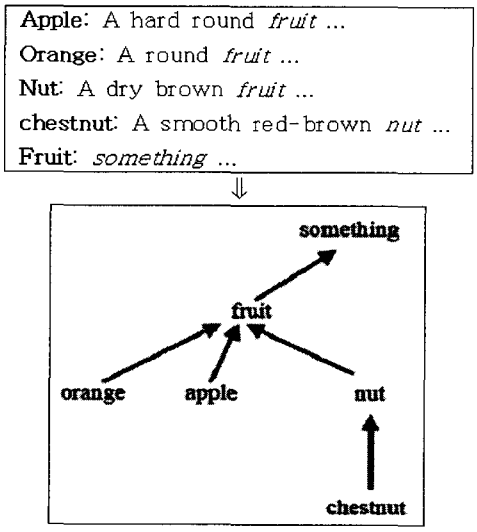
- 원리 1. 기계가독사전은 자연언어처리에 유용한 정보를 포함한다.
- 원리 2. 의미관계는 기계가독사전에서 추출하기가 상대적으로 쉽다.

원리 1은 〈그림 4〉의 "fork"의 정의에서와 같이 자연언어처리 연구에 필수적인 다른 어휘 항목 간의 여러 의미관계를 확인할 수 있으며, 원리 2는 〈그림 5〉와 같이 명사에 대한 정의는 전형적으로 정의항의 명사구의 핵심어가 상위어라는 휴리스틱을 이용하여 자동으로 상위어를 추출할 수 있다는 것이다.

이와 같은 원리로 사전의 정의에서 의미정보를 추출하는데 많은 성과가 있었지만, 자동으로 추출된 의미관계가 완전하지 않으며, 추출된 관계의 품질과 유용성을 평가하는 연구도 거의 없을 뿐만 아니라 정의에서 복잡한 정보를 체계적으로 추출하는 방법이 거의 없다는 지적도 제기되었다.



〈그림 4〉 정의에서 의미정보



〈그림 5〉 정의에서 상위어 추출

3. 표준정의를 이용한 시소러스 확장

3.1 연구개요

오늘날 사용되고 있는 대부분 시소러스의 내용과 구조는 1960년대에 만들어진 시소러스 구축에 관한 국가 및 국제적 지침에 따라 특별한 변화 없이 지금까지 지속되어 왔다. 그러나 시소러스에 수록되는 용어의 수와 관계구조가 증가함에 따라 수록된 용어의 의미는 점차 파악하기 어렵게 되었다.

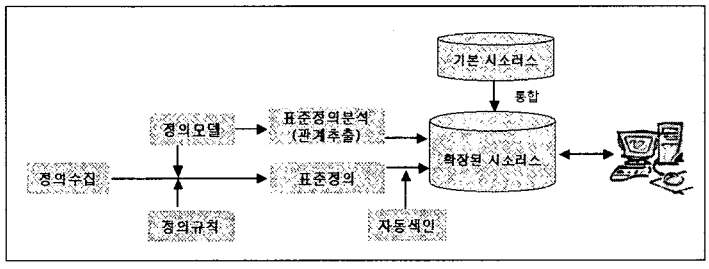
특히 동일한 주제 분야의 시소러스라 하더라도 그 내용이 상당히 다르게 수록되어 있는데, 이러한 문제는 사회과학 분야의 시소러스에서 더욱 심하게 나타나고 있다. 그 이유는 여러 가지가 있겠지만, 기본적으로 시소러스를 구축할 때 수집한 용어에 대한 정확한 의미를 고려하지 않고 내용을 구성한 것이 하나의 큰 이유라 할 수 있겠다.

따라서 이 연구는 시소러스에 용어의 의미를 포함하기 위해 디스크립터에 대한 표준화된 정의를 작성하고 아울러 그 과정에서 만들어진 정의모델의 내용에서 의미관계를 추출하여 이들을 시소러스에 통합하여 확장시키는 것을 목적으로 하였다.

이를 위해, 먼저 기존에 사용되고 있는 시소러스 내용의 일부분을 발췌하여 기본 시소러스로 하고, 이 기본 시소러스에서 디스크립터로 제시된 각 용어에 대해 관련분야의 전문사전이나 텍스트북 및 일반사전 등의 정보원을 이용하여 의미정보를 수집하여 전문용어분야의 이론에서 도출된 정의모델과 정의규칙을 이용하여 표준화된 정의를 작성하였다. 또한 정의모델의 내용에 대해 어휘구문패턴에 근거하여 관계를 추출하고, 앞서 작성된 표준화된 정의와 추출된 관계를 기본 시소러스에 통합하거나 대체하여 확장된 시소러스를 프로토타입으로 구축하였으며, 그 전체과정은 〈그림 6〉과 같이 진행하였다.

3.2 용어의 의미정보수집

이 연구에서 프로토타입의 확장된 시소러스를 만들기 위한 기본 시소러스 데이터는 "KEDI 교육 시소러스"의 31개 주제 중 18개 주제에서 선택한 139개의 디스크립터와 관계를 이용하였다. 선택된 18개 주제 중 17개 주제에서는 교육학 분야의 전체적인 적용가능성을 고려하여 각 주제마다 골고루 모두 74개의 디스크립터를



〈그림 6〉 표준정의를 이용한 확장된 시소러스 구축과정

〈표 1〉 의미정보추출에 사용한 정보원

• 전문사전	- 敎育辭典編纂委員會. 敎育學大辭典(1994) - 南億祐 등 편. 최신 교육학대사전(1990) - 서울대학교 교육연구소 편. 敎育學用語辭典. 전정판(1994) - 현종익, 이학춘. 교육학 용어사전(2002)
• 일반사전	- 엠파스 국어사전(http://kordic.empas.com) - 엠파스 백과사전(http://100.empas.com)
• 텍스트북	- 김정환, 강선보. 교육학개론(1997) - 박준영. 교육학개론(1998) - 임창재. 교육학(2001)

임의로 선정하고, 나머지 한개 주제인 '교육과정'에서는 특정주제에서의 적용가능성을 고려하여 전체 65개 디스크립터 모두를 선정하였다. 다만 나머지 13개 주제는 '도덕교육', '언어교육', '수학교육' 등 특정 교과를 대상으로 한 용어가 대다수여서 배제하였다. 이 과정에서 선정한 디스크립터와 관계는 타당한 것으로 가정하여 기본 시소러스 데이터로 하였다.

일반적으로 전문용어에 대한 의미정보는 그 분야의 정보원에서 추출하거나 주제전문가의 자문을 통해 수집하는데, 이 연구에서는 선정한 139개의 디스크립터에 대한 의미정보를 얻을 정보원으로 〈표 1〉에 수록된 전문사전, 일반사전 및 텍스트북 등을 이용하였으며, 의미정보를 수집한 방법은 다음과 같다.

첫째, 전문사전 네 가지를 조사하여 139개의 디스크립터에 대한 정의가 나타나면 이를 출처와 함께 기록하였다. 이 과정에서 비록 용어가 정확하게 일치하지 않더라도 해당 디스크립터의 번역이나 기술방식의 차이를 고려하여 약간 다르게 기술되어 있는 용어도 동일한 의미로 인정하였다. 그 결과, 한개 이상의 정의가 나타난 디스크립터는 100개로 조사되었다.

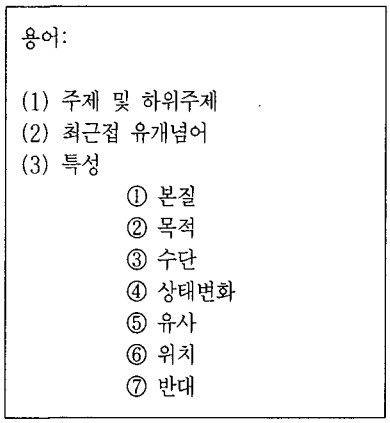
둘째, 전문사전에 정의가 조사되지 않은 나머지 39개 디스크립터는 그것을 구성하고 있는 용어를 분할하여 조사하였다. 이는 대부분의 전문용어가 복합명사로 구성되어 있는 경우가 많기 때문이다. 이 과정에서는 앞서 조사된 100개 디스크립터에 대한 정의와 나머지 일반사전과 텍스트북을 이용하여 조사하였으며, 추가로 30개의 디스크립터에 대한 의미정보를 조사하였다.

전체적으로 기본 시소러스 데이터로 선정한 139개 디스크립터 중 130개(전체의 93.5%)의 디스크립터에 대한 의미정보를 얻었으며, 나머지 9개는 위의 정보원에서 조사되지 않았다. 조사되지 않은 디스크립터는 대부분 그 의미를 해석하기에 포괄적이거나 전문용어로 보기에 다소 문제가 있는 용어로 판단되었는데, 이런 문제는 시소러스에 수록할 용어를 선택하는데 신중을 기해야 할 부분으로 고려된다.

3.3 정의모델과 정의규칙의 설정

일반적으로 전문사전에 수록된 정의는 그 분야의 전문가에 의해 작성되었기 때문에 의미정보를 파악하기에 좋은 정보원으로 여겨지고 있다. 그러나 앞서 조사한 전문사전의 정의를 보면 일정한 규칙이나 기준이 없이 작성되어 내

용이나 형식이 일관성이 없는 것으로 판단되었다. 따라서 본 연구에서 설정한 정의모델은 앞서 언급한 Sager & L'Homme(1994)와 Hudon(1996)이 제시한 정의모델을 근거로 하여 표준정의를 작성하기 위해 〈그림 7〉과 같이 새로운 정의모델을 설정하였다.



〈그림 7〉 새로 설정한 정의모델

이 정의모델에서 항목(1)은 주제와 그 하위주제를 다루는데, 이 연구에서 사용한 'KEDI 교육 Thesaurus'는 교육학 분야의 시소러스로서 31개의 하위주제로 구분하여 구성되어 있기 때문에 '교육학 - 하위주제'형식으로 나타나도록 하였다. 항목(2)는 상위어로서 최근접한 유개념어를 부여하도록 하였다. 이는 시소러스 구조에서 계층관계를 나타내는데 사용할 수 있는데, 특별한 경우를 제외하고는 이 모델로 작성한 모든 표준정의에서 최근접 유개념어가 정의문의 마지막에 제시되도록 하였다. 마지막으로 항목(3)은 정의항에서 피정의항과 동일한 수준의 다른 디스크립터와 구별되는 특성을 제시한 것으로 7가지를 설정하였다. 설정된 특성은 130개 디스크립터의 의미정보에 대해 지금까지 제시된 모든 특성들을 적용하여 해당되는 것만 선택하여 구분한 것으로 교육학 분야에 타당성이 있을 것으로 판단된다.

정의규칙은 수집한 의미정보를 앞서 설계한 정의모델에 의거하여 분석한 다음, 이해하기 쉽고 일관성 있는 표준정의로 작성하기 위한 것이다. 이 연구에서 설정한 정의규칙은 대부분 앞장에서 언급한 표준에서 제시된 것을 기초로 어휘의미적 규칙과 형태구문적 규칙으로 구분하여 다음과 같은 내용으로 구성하였다.

〈어휘의미적 규칙〉

규칙 1. 정의항에서는 피정의항의 상위개념어을 반드시 포함한다.

규칙 2. 정의항에서는 피정의항에 이용된 용어를 반복하여 사용하지 않는다.

규칙 3. 정의항의 상위개념어가 다른 곳에서 정의될 때, 그것의 특성을 다시 제시하지 않는다.

규칙 4. 상위개념을 파악할 수 없는 개념은 상위개념어을 부여하지 않는다.

규칙 5. 상위어가 시소러스에 없을 때는 새로운 용어를 부여하며, 범주를 나타내는 용어는 그 개념을 고정하는데 사용하나 상위어로 인정하지 않는다.

규칙 6. 정의항에서 피정의항의 동의어를 사용하지 않는다.

규칙 7. 필요에 따라 정의항에 피정의항의 반대개념을 부여할 수 있다.

〈형태구문적 규칙〉

규칙 8. 정의는 문법적으로 완전한 하나의

문장으로 구성된 명사구로 한다.

규칙 9. 동사는 현재시제로 하고 능동태가 되어야 한다.

3.4 표준정의작성

시소러스에서 표준정의의 기능은 디스크립터로 표현된 개념에 대해 충분한 의미정보를 제공하여 이용자가 디스크립터를 선택하여 이용할 때 다른 디스크립터와 구별을 명확히 하는 것이다. 이 연구에서는 앞서 설정한 정의모델과 정의규칙에 따라 139개 디스크립터 중 정보원을 통해 표준정의를 작성하는데 필요한 의미정보가 조사된 130개의 디스크립터에 대해 다음의 사례와 유사한 패턴으로 표준정의를 작성하였다.

〈사례〉 '통합교육과정'(Integrated Curriculum)의 표준정의

이 용어에 대한 의미정보는 정보원에서 다음과 같이 세 개가 조사되었다.

- 정의 1: 현대사회에서 개인이나 사회가 당면하는 문제들을 해결하기 위해서, 필요로 하는 지식을 통합하여 응용할 수 있는 통합적 지식을 획득하는 교육과정으로서, 학문중심 교육과정에서 변형된 교육과정이다.
- 정의 2: 교과영역에 구애됨이 없이 이들을 횡단하여 일정한 기준에 따라 학습내용 및 경험을 선정, 조직하려는 교육과정
- 정의 3: 통합과정(integrated program)이라고도 한다. 통합과정이라는 이름 밑에 상관과정, 광역과정, 융합과정, 문제중심과정 혹은 경험과정으로 다시 분류할 수 있다.

이 용어는 교육학 분야에서 주제가 '교육과정'에 포함된 것이므로 정의모델의 항목 중 주제분야는 '교육학 - 교육과정'이다. 또한 조사한 정의를 분석한 결과 유개념어는 '교육과정'이며, 특성형태는 다음의 네 가지로 나타났다. 특히 조사된 기존의 정의에서 '교육과정'에 대한 정의에서와 동일한 특성이 반복되어 나타나지만, 상위개념의 특성을 상속하는 부분은 여기에 포함시키지 않았다.

- 특성①(본질): 교과영역에 구애됨이 없이; 필요한 지식을 통합하여 응용할 수 있는 통합적 지식을 획득하는
- 특성②(목적/기능): 현대사회에서 당면하는 문제들을 해결하기 위해
- 특성④(상태변화): 학문중심교육과정에서 변형된; 상관과정, 광역과정, 융합과정, 문제중심과정, 경험과정으로 분류
- 특성⑤(유사): 통합과정

이들 특성형태 중에서 특성①과 특성④에 해당하는 내용이 두 가지가 나타났으며, 이를 토대로 '통합교육과정'의 표준정의를 다음과 같이 작성하였다.

- 표준정의: 학문중심교육과정과 달리 현대사회에서 당면하는 문제들을 해결하기 위해 필요로 하는 지식을 통합하여 응용할 수 있는 통합적 지식을 획득하는 교육과정

이상과 같이 모든 디스크립터에 대한 표준정의를 작성한 결과, 다음의 사항을 확인하였다.

첫째, 정의모델을 통해 계층관계로서 용어의 상위어인 최근접 유개념어를 확실하게 파악할 수 있었다. 전체 130개의 용어 중 유개념어가 있는 디스크립터는 125개로 나타났으며, 나머지 다섯 개의 용어는 최상위어로 파악되어 유개념어가 나타나지 않은 것으로 조사되었다. 둘째, 각 디스크립터의 의미정보를 분석하여 도출한 특성은 대부분 한 두개 정도로 파악되었는데, 특별히 몇몇 주요한 용어는 세 개 이상의 특성을 가지는 것으로 조사되었다.

아울러 사용된 특성분포를 보면, 특성①(본질)이 가장 많이 이용되었으며, 그 다음으로 특성②(목적/기능)이 이용된 것으로 나타났다. 특히 특성⑤(유사)는 나머지 특성보다 상대적으로 많이 이용된 것으로 나타났는데, 이것은 유사어를 이용하여 용어의 의미를 전달하는 것이 효과적임을 보여준다고 할 수 있다. 이러한 특성형태로 구분한 것은 조사된 의미정보에서 다음의 문장형태로 나타나는 것을 근거로 한 것이다.

- 특성①(본질): "...을 하는"
- 특성②(목적/기능): "...을 위한", "...하도록"
- 특성③(수단/도구): "...하여", "...를 통해"
- 특성④(상태변화): "...에서 생긴"
- 특성⑤(유사): "...라고 하는"
- 특성⑥(위치): "...에서"
- 특성⑦(반대): "...와 다른", "...와 구분하는", "...에 대립하여", "...이 아니라"

3.5 표준정의를 통한 관계추출

표준정의를 통해 의미관계를 추출하는 방법은 표준정의를 작성할 때 사용한 정의모델의 내용을 근거로 하였는데, 추출한 관계는 기존의 시소러스와 통합을 고려하여 BT, RT, NT로 한정하였다. 표준정의에서 관계를 추출하는 방법은 먼저 표준정의를 구문 분석한 다음, 미리 선정한 관련 있는 어휘구문패턴으로 확인하여 일치하는 부분을 선정하는 것으로 하였다. 그러나 자연어로 구성된 표준정의의 형식과 내용이 매우 다양하기 때문에 자동으로 구축하는데 어려움이 많아 이 연구에서는 정의모델의 항목을 근거로 하여 수작업으로 다음과 같이 관계를 추출하였다.

먼저 정의모델의 항목 중 최근접 유개념어에 해당되는 용어는 기본적으로 상위어가 되므로 BT에 해당하는 것으로 하였다. 이에 따라 130개의 디스크립터 중 125개의 디스크립터에 대한 상위어가 추출되었으며, 나머지 5개는 최상위어이어서 추출되지 않았다. 특히 대부분 시소러스에서 상위어의 계층수준이 불분명한 반면, 이 상위어는 디스크립터에 대해 가장 인접한 상위어가 되기 때문에 용어의 계층수준을 명확하게 하는 것으로 파악되어 의미 있는 결과로 볼 수 있다.

또한 정의모델의 특성항목에서 제시된 내용을 통해 수작업으로 다음의 사례와 유사한 방법으로 RT와 NT에 해당하는 관계를 추출하였다.

〈사례 1〉 정의모델 중 유사특성 및 반대특성에서의 관계추출

유사특성에 나타나는 용어는 모두 RT에 해당하는 것으로 하고, 반대특성에서 "...와 대조적으로", "...와 다른", "...와 구분하는", "...에 대립하여", "...이 아니라"라는 패턴으로 나타

나는 명사는 RT에 해당하는 것으로 하였으며, 그 사례는 〈그림 8〉과 같다.

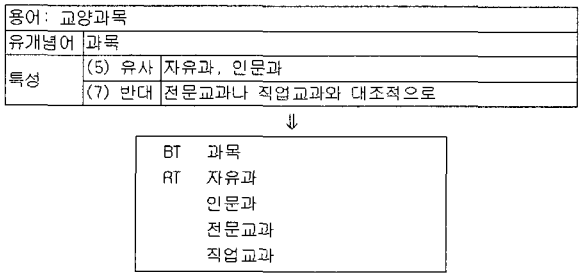
〈사례 2〉 정의모델 중 본질특성에서의 관계추출

본질특성에서 "... 포함하는", "... 조직하는", "... 구성하는" 이라는 패턴으로 나타나는 명사는 NT에 해당하는 것으로 하였으며, 그 사례는 〈그림 9〉와 같다.

〈사례 3〉 정의모델 중 상태변화특성에서의 관계추출

상태변화특성에서 "... 으로 이행하는" 이라는 패턴으로 나타나는 명사는 RT에 해당하는 것으로 하였으며, 그 사례는 〈그림 10〉과 같다.

이처럼 정의모델의 특성항목 중 특성①(본질), 특성④(상태변화), 특성⑤(유사), 특성⑦(반대)에서 주로 의미관계가 추출되었으며, 나머지 특성항목들은 대부분 정의의 의미를 명확하게 하기 위해 기술한 것으로 의미관계가 추출되지 않았다. 이상의 방법으로 정의모델의 내용을 통해 전체 130개의 용어 중 125개의 BT, 80개의 RT, 57개의 NT 관계가 새로이 추출되었다.

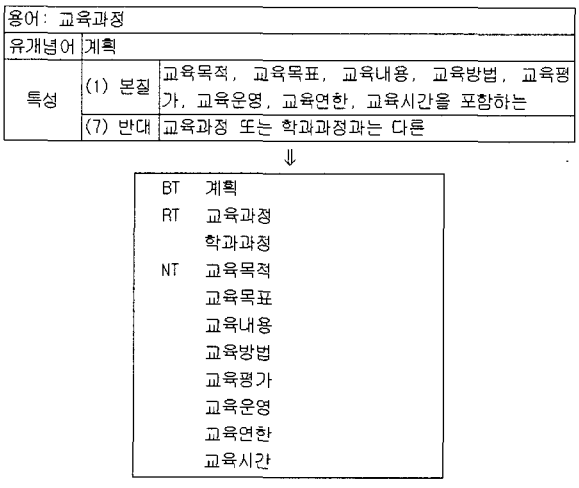


용어: 교양과목		
유개념어	과목	
특성	(5) 유사	자유과, 인문과
	(7) 반대	전문교과나 직업교과와 대조적으로

⇓

BT	과목
RT	자유과
	인문과
	전문교과
	직업교과

〈그림 8〉 유사특성 및 반대특성에서 나타나는 관계



용어: 교육과정		
유개념어	계획	
특성	(1) 본질	교육목적, 교육목표, 교육내용, 교육방법, 교육평가, 교육운영, 교육연한, 교육시간을 포함하는
	(7) 반대	교육과정 또는 학과과정과는 다른

⇓

BT	계획
RT	교육과정
	학과과정
NT	교육목적
	교육목표
	교육내용
	교육방법
	교육평가
	교육운영
	교육연한
	교육시간

〈그림 9〉 본질특성에서 나타나는 관계

용어: 융합교육과정		
유개념어	교육과정	
특성	(4) 상태변화	상관교육과정에서 광역교육과정으로 이행하는 과정에서 생긴

⇓

BT	교육과정
RT	상관교육과정
	광역교육과정

〈그림 10〉 상태변화특성에서 나타나는 관계

3.6 시소러스 확장

기본 시소러스에 새로 작성한 표준정의와 추출된 관계를 통합하여 확장된 시소러스를 구축하는 과정은 다음과 같다. 먼저 새로 작성한 표준정의는 제시어 'DF'를 부여하여 기본 시소러스에 그대로 통합하였고, 정의모델의 내용에서 추출한 관계를 기본 시소러스에 추가로 통합하거나 대체하여 시소러스를 확장하였다.

이러한 과정을 통해 구축된 확장된 시소러스의 구조와 내용은 〈그림 11〉의 사례와 같이 변화되었다. 여기서 굵은 글씨체로 표시된 것은 변화가 있는 부분이 되며, 제시어 중 SB는 주제분야, DF는 표준정의를 나타내고 나머지 관계는 기존의 시소러스와 동일하다.

특히, 새로이 추가된 관계 중 BT는 정의모델에서 조사된 유개념어가 해당 용어의 최근접 상위어가 되기 때문에, 기존 시소러스에서 계층구조의 깊이를 고려하지 않고 부여한 BT보다 계층수준을 더욱 명확히 하여 상당한 의의

기존의 시소러스	⇒	통합된 시소러스
교육목표		교육목표
BT 목표		**SB 교육학 - 교육과정**
RT 교육철학		**DF 교육목적과 교육성과를...**
심체적목표		**BT 교육과정**
인지적목표		RT 교육철학
정의적목표		심체적목표
행동목표		인지적목표
		정의적목표
		행동목표
		교육목적
		교육성과
		교육내용
나선형 교육과정		나선형 교육과정
BT 교육과정		**SB 교육학 - 교육과정**
RT 교육과정개발		**DF 지식구조를 가르치기 위해...**
교육과정기획		BT 교육과정
		RT 교육과정개발
		교육과정기획
		지식구조

〈그림 11〉 표준정의와 추출된 관계를 통합한 시소러스의 사례

〈표 2〉 BT의 적용 양상

- BT가 변경된 용어	59개(47.2%)
- BT가 새로 부여된 용어	24개(19.2%)
- BT가 변경되지 않은 용어	42개(33.6%)

가 있는 것으로 볼 수 있다. 이 연구에서는 표준정의를 작성한 130개의 디스크립터 중 최상위 용어인 다섯 개를 제외한 125개의 용어 중에서 BT의 적용양상은 〈표 2〉와 같이 나타났다. 이 결과를 볼 때 기존 시소러스의 BT가 변경되거나 새로 부여된 BT가 전체 66.4%정도가 되어 시소러스의 계층수준에 있어서 상당한 변화를 보여주었으며, 그대로 적용되는 것은 33.6%정도로 기존 시소러스의 BT가 상대적으로 타당성이 낮음을 알 수 있다.

4. 결 론

색인과 탐색과정에서 선정한 용어간의 일치여부에 근거하는 정보검색과정에서 시소러스는 자연언어의 특성상 나타나는 용어와 개념간의 불일치를 해소하여 개념기반검색을 가능하게 한다. 이러한 역할을 위해 시소러스는 기본적으로 관계구조나 범위주기를 이용하여 용어의 의미뿐만 아니라 용어간의 관계를 명확하게 제시하여야 한다.

그러나 용어의 의미를 제시하는 문제는 기존의 관계구조만으로는 한계를 보여 왔다. 즉, 기본관계나 범위주기만으로는 용어가 지니는 본질적인 속성을 이해하기 어려울 뿐만 아니라 그 범위를 한정하기 어려운 면이 있기 때문에 정보검색과정에서 용어선택시 일관성을 유지하고 용어의 의미와 용어 및 개념간의 관계를 분명히 파악하기 어려운 면이 있었다. 이에 일부 시소러스는 사전의 정의를 포함하기도 하였으나, 수록한 정의의 내용이나 형식이 만족스럽지 않아 명확한 의미를 전달하는데 문제가 있는 것으로 지적되었다.

이에 본 연구는 용어에 대한 표준화된 정의를 작성하고, 아울러 그 내용을 통해 관계를 추출하여 기존의 시소러스에 통합하여 확장시킴으로서 그 유용성을 높이고자 하였다. 이러한 목적에 의거하여, 실제적으로 'KEDI 교육 Thesaurus'의 31개 하위주제 중 18개 하위주제에서 선정한 139개의 디스크립터와 관계를 기본 시소러스로 하여 다음의 단계로 진행하여 프로토타입의 확장된 시소러스를 구축하였다.

(1) 관련 정보원에서 의미정보의 수집과 기록
(2) 수집한 의미정보를 정의모델에 따라 분석 및 기록
(3) 분석된 정의모델의 내용을 정의규칙에 따라 표준정의로 작성
(4) 정의모델의 데이터 요소에서 의미관계 추출
(5) 표준정의와 추출된 의미관계를 기존 시소러스에 통합

그 결과, 시소러스에서 각 디스크립터에 대한 의미로서 표준정의가 모두 부가되어 용어의 의미를 명확하게 파악할 수 있게 하였으며, 새로 추출된 의미관계를 통합하여 기존의 시소러스가 더욱 확장되는 것으로 나타났다. 특히 추출된 의미관계에서 125개의 상위어 중 59개(47.2%)가 기존의 상위어를 대체하였으며, 24개(19.2%)가 새로 생성되어 계층수준을 더욱 명확하게 하는 것으로 나타났다. 이는 지금까지 기존의 시소러스에서 제시된 상위어가 계층수준이 일

정하지 않거나 2개 이상의 상위어가 제시되어 명확한 최근접 상위어를 파악하기 힘든 것과 대조적으로 명확한 계층수준을 제시할 수 있음을 보여 주었다.

이러한 연구과정을 통해, 표준정의를 이용하여 확장된 시소러스는 정보검색과정에서 용어의 의미를 정확하게 파악하여 더욱 분별 있게 용어를 선택할 수 있게 하며, 시소러스의 관계구조의 확장뿐만 아니라 더욱 적절한 계층수준을 제시하여 변화하는 정보환경에 적응하는 시소러스의 사례를 보여주었다. 나아가 이를 더욱 발전시키기 위해서는 다양한 주제 분야의 용어의 표준정의를 작성하기 위한 적절한 정의 모델과 정의규칙의 개발을 통해 자동으로 표준정의가 작성될 수 있는 방법과 어휘구문패턴을 적용하여 자동으로 관계를 추출하는 방법이 향후 연구되어야 할 것이다.

참 고 문 헌

이병근. 1992. 辭典 定義의 類型과 原則. 『새국어생활』, 1(1): 2-21.

Aitchison, J. 1994. *Words in the Mind: an Introduction to the Mental Lexicon.* Oxford: Blackwell.

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

Aitchison, Jean, Alan Gilchrist, and David Bawden. 2000. *Thesaurus Construction and Use: a Practical Manual.* 4th ed. London: Aslib.

Amsler, R. A. 1980. *The Structure of the Merriam-Webster Pocket Dictionary.* Doctoral Dissertation. TR-164, University of Texas, Austin.

Buchan, R.L. 1989. "Intertwining thesauri and dictionaries." *Information Services & Use.* 9: 171-175.

Chodorow, Martin S., Roy Byrd, and George Heidorn. 1985. "Extracting semantic hierarchies from a large on-line dictionary." In *Proceedings of the 23th Annual Meeting of the Association for Computational Linguistics*, 299-304.

Condamines, A. and J. Rebeyrolle. 2001. "Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base(CTKB): method and results." In D. Bourigault, C. Jacquemin, and M.-C. L'Homme (eds.). *Recent Advances in Computational Terminology.* Amsterdam: John Benjamins Publishing company, 127-148.

Dahlberg, I. 1981. "Conceptual definitions for interconcept." *International Classification*, 8: 16-22.

Dahlberg, I. 1989. "Concept and Definition Theory." In *Classification Theory in the Computer Age: Conversations Across the Disciplines. Proceedings form the Conference, November 18-19, 1988.* Albany, New York: Rockefeller College Press, University of Albany, State University of New York, 12-24.

Felber, H. 1984. *Terminology manual.* Paris: UNESCO.

Hindle, Donald. 1990. "Noun classification from predicate-argument structures." In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 268-275.

Hudon, Michèle. 1996. "Preparing terminological definitions for indexing and retrieval thesauri: a model." In Rebecca Green (ed), *Knowledge organization and change. Proceedings of the fourth International ISKO Conference, Washington, DC, 15-18 July 1996.* Frankfurt/Main: Indeks Verlag, 363-369.

Hudon, Michèle. 1998. *An Assessment of the Usefulness of Standardized Definitions in a Thesaurus Through Interindexer Terminological Consistency Measurements.* Ph.D. diss., University of Toronto.

Ide, Nancy and Jean Véronis. 1993. "Extracting knowledge-bases from machine-readable dictionaries: have we wasted our time?." In *Proceedings KB&KB'93 Workshop*, 257-266.

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

International Organization for Standardization. 1987. *ISO 704-1987: Principles and Methods of terminology.* Geneva: ISO.

Markowitz, J., T. Ahlswed and M. Evens. 1986. "Semantically significant patterns in dictionary definitions." In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, 112-119.

McNaught, J. 1983. "The generation of term definitions from an on-line terminological thesaurus." In *First Conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the Conference, 1-2 September 1983, Pisa, Italy.* Morristown, JN: Association for Computational Linguistics, 90-95.

Michiels, A. and J. Noël. 1982. "Approaches to thesaurus production." In *Proceedings of the 9th International Conference on Computational Linguistics(COLING-82)*, 227-232.

Mooers, C. N. 1985. "The indexing language of an information retrieval system." In *Theory of Subject Analysis: a Sourcebooke.* Littleton, CO: Libraries Unlim-

ited, 247-261.

National Information Standards Organization. 1994. *ANSI/NISO Z39.19-1993: Guidelines for the Construction, Format, and Management of Monolingual Thesauri.* Bethesda, MD: NISO.

Ndi-Kimbi, Augustin. 1994. "Guidelines for terminological definitions: the adherence to and deviation from existing rules in BS/ISO 2382 - Data Processing and Information Technology Vocabulary." *Terminology*, 1: 327-350.

Nkwenty-Azeh, B. 1994. "The use of thesaural facets and definitions for the representation of knowledge structures." In Hanne Albrechtsen and Susanne Oernager (eds.), *Knowledge Organization and Quality Management: Proceedings of the Third International ISKO Conference, 20-24 June, 1994, Copenhagen, Denmark.* Frankfurt: Indeks Verlag, 374-381.

Sager, J.C. 1982. "Terminological thesaurus: a more appropriate designation or a deprecated synonym?." *Social Science Information Studies*, 2:211-214.

Sager, J.C. 1990. *A Practical Course in Terminology Processing.* Amsterdam: J. Benjamins.

Sager, J.C., and M.C. L'Homme. 1994. "A model for the definition of concepts: rules for analytical definitions in terminological databases." *Terminology*, 1:351-373.

Sager, Juan C. and Augustin Ndi-Kimbi. 1995. "The conceptual structure of terminological definitions and their linguistic realization: A report on re- search in progress." *Terminology*, 2(1): 61-81.

Sager, J.C., H.L. Somers, and J. McNaught. 1982. "Thesaurus integration in the social sciences. Part III: Guidelines for the integration of thesauri." *International Classification*, 9: 64-70.

Soergel, D. 1974. *Indexing Languages and Thesauri: Construction and Maintenance.* Los Angeles: Melville.

Soergel, D. 1985. *Organizing Information: Principles of Data Base and Retrieval Systems.* Orlando: Academic Press.

Svenonius, E.. 1990. "Design of controlled vocabularies." In *Encyclopedia of Library and Information Science.* New York: Marcel Dekker, 82-109.

Svenonius, E.. 1997. "Definitional approaches in the design of classification and thesauri and their implications for retrieval and for automatic classification." In: *Knowledge Organization for Information Retrieval. Proceedings of the 6th International Study Conference on Classification Research, University College London, 16-18 June 1997.* The Hague, Netherlands: International Federation for Information and Documentation, 12-16.