新聞 시소러스의 평가에 관한 硏究

— 『新聞記事 綜合시소러스』를 中心으로 —

A Study on the Evaluation of Newspaper Thesaurus

李 仁 愛 (In-Ae Lee)

□ 목 차 □

1.서론
2.『신문기사 종합시소러스』의 평가방법
2.1 색인작업
2.2 시소러스와 신문용어 수집 리스트와의 대조
3.『신문기사 종합시소러스』의 평가결과
3.1 색인작업의 결과
3.2 신문용어 수집 리스트와의 대조 결과
4.고찰
4.1 주제개념에 대한 표현력
4.2 주제분야에 대한 포괄성
5.결론

초 록

본 연구는 신문 시소러스의 평가에 관한 연구의 일환으로서, 『신문기사 종합시소러스』의 경제 · 산업분야를 대상으로 시소러스의 표현력과 포괄성을 평가하였다. 평가는 시소러스를 사용하여 신문기사에 대해서 색인작업을 하고, 신문기사와 용어사전에서 수집한 신문용어를 시소러스와 대조하는 두가지의 방법으로 하였다. 평가를 통하여, 신문 시소러스의 구축과 이용에 있어서 중요한 문제로 주제개념의 특정성, 복합어의 분리, 디스크립터와 도입어간의 우선관계, 고유명사의 수록방법 그리고 주제분야간의 용어배분 문제가 고찰되었다.

ABSTRACT

This study evaluates representability and comprehensivity of the Theasurus in the economics and industry fields of the "General Thesaurus of Newspaper Articles." The methods used in the study were, first, indexing of the pages covering economics and industry articles using the Thesaurus and second, comparing the Thesaurus terms with the words collected from the newspapers articles and glossaries. The study clarifies the following problems which might occur in the construction and use of newspaper thesaurus: specificity of the subject concepts, separation of component concept, preference relationship between descriptors and entry terms, the methods of recording of proper nouns and allocation of terms among the subject areas concern.

---

* 본고는 1994년 日本 圖書館情報大學 圖書館情報學硏究科 석사학위 논문의 요약임.

■ 논문접수일 : 1995년 4월 25일

1. 서 론

최근, 신문 데이터베이스의 구축은 컴퓨터에 의한 신문의 편집과 제작을 행하는 CTS (Computerized Typesetting System)의 개발과 더불어 커다란 진전(松尾光,神尾達夫, 1993)을 보이고 있으며, 사회에서 일어난 일을 망라적으로 전하는 速報的인 정보로서, 신문정보의 유효성을 한층 높이게 되었다.

이러한 신문 데이터베이스 중에는, 작성과 검색의 효율화를 도모하기 위하여, 신문기사의 주제가 되는 용어의 의미적 관계를 통제한 용어집(阿部哲也, 1985)인 신문 시소러스를 사용하고 있는 것도 있다. 시소러스는 동의어와 유의어의 통제및 관련된 의미를 갖는 용어간의 관계를 통제하여, 색인부여에 있어서는 색인자가 문헌의 주제내용을 일관된 방법으로 표현할 수 있도록 하며, 검색에 있어서는 어떤 주제에 대해서 포괄적인 검색이 용이하게 될 수 있도록 하는 기능을 갖고 있다.(Lancaster 1976)

그러나, 시소러스의 작성은 방대한 시간과 인력이 필요하며, 통제하는 용어가 변화하여 가기 때문에, 유지와 관리에도 상당한 인력을 필요로 한다. 또한, 신문 데이터베이스는 속보성이 있는 신문정보를 제공하기 때문에 시간적 지연(time lag)이 가능한 한 짧은 것이 요구되지만, 매일의 방대한 신문기사를 색인자에 의해 색인작업을 행하는 것은 상당히 인력과 시간을 필요로 하기 때문에(神尾達夫 1989) 자동 색인작업(automatic indexing)을 행하고 있는 신문사가 많아지고 있다. 확실히, 자동 색인작업은 컴퓨터가 모든 정보를 일정의 절차를 거쳐 처리하기 때문에, 방대한 기사를 처리하기에는 사람에 의한 작업보다 단시간에 처리가 가능하고, 인건비등의 비용을 절약하며, 색인자의 개인차에 의한 색인작업의 분산도 없다는 장점이 있다(菊池敏典 1992).

그러나, 현재의 자동 색인작업은 주제분석을 행하지 않기 때문에, 기사중에 표현되어 있지 않은 중요한 주제개념이 색인되지 않으며, 기사의 의미를 판단하지 않기 때문에 同音異意語와 같은 용어를 구분할 수 없다는 등의 문제가 있다(神尾達夫 1989). 더우기, 키워드가 기계적으로 부여되기 때문에 주제와 무관계한 말이 키워드로 추출되어 버리는 일도 있으며, 추출된 다량의 키워드는 데이터베이스의 관리에 있어서 부담이 된다(廣木守雄,阿部哲也 1991). 이 때문에, 현단계의 신문 데이터베이스의 자동 색인작업은 완전히 자동화 되고 있지는 않고, 컴퓨터에 의해 부여된 키워드의 精度를 높이도록 다수의 사람에 의해 손질을 가하는 신문사도 있다(羽原隆司 1992).

이러한 자동 색인작업의 문제에 대처하기 위하여도, 신문 시소러스는 활용될 수 있다. 예를들면, 자동 색인작업의 과정에 있어서, 추출된 용어를 키워드로서 판정할 때에 기준으로 되는 기본 사전으로서, 시소러스를 활용하는 것에 의해(廣木守雄,阿部哲也 1991) 동의어를 정리하거나 주제개념과 무관한 말을 제외하는 것 등으로, 키워드의 精度를 높일 수 있다. 이와같이, 시소러스는 자동 색인작업에 있어서도 유효한 수단이 될 수 있지만, 현시점에서는 신문 데이터베이스의 색인작업이 사람에 의존하는 부분이 남겨져 있다는 점에서도, 시소러스는 필요하다고 할 수 있다. 특히, 신문기사는 사회현상을 대상으로 하기 때문에 시간의 경과

에 따른 변동이 크고, 하나의 기사만이 아니고 몇개의 기사를 합하므로서 정보로서의 가치를 갖게 되는 경우가 많다. 그 때문에, 신문정보의 검색에 있어서는, 목적으로 하는 기사 1,2건을 찾는 경우보다도, 어떤 테마에 관계하는 기사를 빠짐없이 검색하여 테마 전체의 동향을 파악하거나 경향을 분석하는 것이 필요하게 된다. 일련의 관련성있는 기사를 합해서 얻기 위해서는 그것들의 기사에 공통의 키워드가 부여되어 있는 것이 필요하기 때문에(神尾達夫, 1989), 일관성이 있는 키워드를 부여하기 위해서 시소러스가 필요하다고 할 수 있다.

이와같이, 신문 데이터베이스에 있어서 시소러스는 중요한 역할을 담당한다고 할 수 있지만, 신문 시소러스에 대해 충분히 연구되어 있다고 할 수 없는 실정이다. 이에, 본 연구에서는, 신문 시소러스의 평가를 통하여 신문 시소러스의 특성과 문제점을 밝히기 위하여, 한국에서 처음으로 작성된 신문 시소러스인 「신문기사 종합시소러스」(한국언론연구원, 한국조사기자회 1992)를 사례로서 평가를 하였다.

시소러스의 평가에 관한 연구에는 다양한 평가기준이 있다. 그 중에서, 精度와 再現率을 기준으로 평가하는 것은(Kristensen 1990) Lancaster(1979)에 의해서도 지적되었듯이, 검색 시스템을 통하여 그 평가를 하기 때문에 시소러스의 성능이외에 색인부여, 검색전략, 이용자와 시스템간의 상호작용 등의 요인이 검색결과에 영향을 미치므로, 시소러스 자체가 검색에 있어서 정도와 재현율을 높이는 것에 어느 정도 유효한가를 측정하는 것은 어렵다고 할 수 있다. 또한 시소러스를 형태적인 측면에서 평가하는 것은(Sager 1981;1982), 색인과 검색의 도구로서, 시소러스의 성능을 결정적으로 제어하는 것은 아니기 때문에 피상적인 평가에 머무를 수가 있다.

이에, 본 연구에서는 시소러스의 디스크립터가 주제개념을 어느정도 적절히 표현할 수 있는가 하는 表現力을 기준으로 평가하기 위해, 시소러스를 사용하여 신문기사의 색인작업을 하였다. 또한 시소러스가 다루고 있는 주제분야의 용어를 어느 정도 포괄할 수 있는가 하는 包括性을 평가하기 위해 신문용어를 수집하여 그 용어 리스트와 대조를 하였다.

그런데, 시소러스에 관하여 사용되고 있는 용어의 의미가 사용자에 따라서 다른 경우가 있기 때문에, 본고에서는 以下와 같이 용어의 의미를 통일하였다. 이를 위해, 색인작성과 시소러스작성을 위한 국제규격인 ISO 2788에 의거한 일본공업규격인 JIS X0901의 정의를 기준으로 하였다.

JIS X0901(日本工業標準調査會 1991)의 정의에 의하면, 색인어는 어떤 개념을 명사 또는 명사구의 형태로 적절히 표현한 것이다. 시소러스에 있어서, 색인어는 우선어인가 비우선어인가가 명시된다. 우선어는 어떤 개념을 표현하는 색인작업에 일관하여 사용되는 말이고, 디스크립터 혹은 키워드라고도 불리지만, 본 연구에서는 〔디스크립터〕를 사용하였다. 또한, 비우선어는 우선어의 동의어 혹은 준동의어로서, 문헌에는 부여되어 있지 않으나, 시소러스에는 표목어로서 취급되어 있는 용어이며, 비디스크립터로 불리지만, 본 연구에서는 USE라는 지시에 의해서 적절한 우선어로 이용자를 안내한다는 의미로서 〔導入語〕를 쓰기로 했다. 따라서, 〔색인어〕는 디스크립터와 도입어의 총

칭으로서 사용한다. 동의어 준동의어에 대해서는 〔同意·類意語〕라는 용어를 쓴다.

2. 『신문기사 종합시소러스』의 평가방법

2.1 색인작업

신문기사의 주제개념에 대한 『신문기사 종합시소러스』의 表現力을 평가하기 위해, 신문기사에 대해서 『신문기사 종합시소러스』를 사용하여 색인작업을 하였다. 대상으로한 신문은, 『신문기사 종합시소러스』가 대상으로 하고 있는 일반지 중에서,『동아일보』와 『조선일보』이다. 이 두가지 신문의 경제·산업분야를 취급하는 페이지에서, 1992년 5월부터 1993년 4월까지 1년간의 기사를 추출했다.

신문기사의 추출방법은, 먼저 각각의 신문에서 26일분을 무작위로 추출하여 얻어진 모든 기사에 일련의 번호를 붙였다. 그 경우, 광고는 제외하고, 표제가 있는 기사를 하나의 기사로서 판단하였다. 이렇게 하여 『동아일보』에서 130건, 『조선일보』에서 136건, 합계 266건의 기사가 색인작업의 대상으로 얻어졌다. 이러한 방법으로 얻어진 266건의 신문기사에 대해서, 주제분석을 하고, 주제개념을 추출했다.

신문이 기사의 중요한 내용을 선두에 기술하는 특징을 갖기 때문에, 주제개념의 추출은 표제를 중심으로 하고, 그 이외의 중요한 개념을 본문으로 부터 추출하였다. 신문기사는 도서와 잡지기사와는 달리 사회현상을 대상으로 하고 있기 때문에, 신문기사의 중요한 주제요소인 5W1H를 표현하는 개념이 추출되도록 하였다.

또한, 신문기사의 주제개념을 추출할 때, 색인작업의 분산이 생기지 않도록, 평균 100자에 1語의 비율로 주제개념을 추출하였다. 신문기사의 길이는 신문과 기사의 유형에 따라서 다르지만, 이번 대상의 신문기사는 평균 400-500자이였다. 이들 기사에 대해, 『신문기사 종합시소러스』 중에서 각 주제개념을 표현하는 디스크립터를 부여하였다. 단, 하나의 주제개념에 대해서 복수의 디스크립터를 합성하여 표현한 경우도 있다. 예를들면 기사의 길이가 400-500자인 경우, 주제개념의 수는 추출될 때 4-5개이지만, 그 이상의 디스크립터가 부여된 경우도 있다.

2.2 시소러스와 신문용어 수집 리스트와의 대조

신문기사의 주제분야에 대한 包括性을 평가하기 위해, 신문기사에서 쓰여지고 있는 경제·산업분야의 용어를 수집하고, 그 용어 리스트와 『신문기사 종합시소러스』의 경제·산업분야에 수록되어 있는 용어와 대조를 하였다.

신문용어의 수집은, 최신의 용어를 수집하기 위해서 신문기사에서 주제어를 추출하였으며, 보다 망라적으로 수집하기 위해서 신문 용어사전과 분류표에서도 수집을 하였다. 신문기사에서는, 『동아일보』의 1992년 1월부터 12월까지의 1년간의 경제·산업분야를 취급하는 페이지에서, 무작위로 12주간분을 선택하여, 그 12주간분의 기사에서 주제어를 추출하였다. 또한 신문사에서 출판된 3개의 신문 용어사전인 『經濟新語辭典』(매일경제신문사

1991), 『現代時事用語辭典』(동아일보사 1992), 『現代知識情報事典』(중앙일보사 1992)와 한국언론연구원과 한국조사기자회에서 편집된 『記事資料標準分類表』(한국언론연구원,한국조사기자회 1991)의 4개의 자료에 대해서, 각 자료간의 수록용어의 중복을 조사하고, 셋 이상의 자료에 수록되어 있는 용어를 수집하였다. 이렇게 하여 신문기사에서 1149語, 신문 용어사전과 분류표에서는 293語가 수집되어, 전체 1442語가 얻어졌다. 이러한 방법으로 수집된 1442어를 음순으로 배열한 신문용어 리스트와 『신문기사 종합시소러스』를 대조하고, 일치하는 용어와 일치하지 않는 〔누락〕의 용어로 나누었다. 이렇게 일치와 누락으로 나뉘어진 수집용어는, 다시 보통명사와 고유명사로 나뉘어졌다. 보통명사에 대해서는 『신문기사 종합시소러스』의 경제 · 산업분야의 카테고리에 의해서 분류를 하였다.

3. 『신문기사 종합시소러스』의 평가결과

3.1 색인작업의 결과

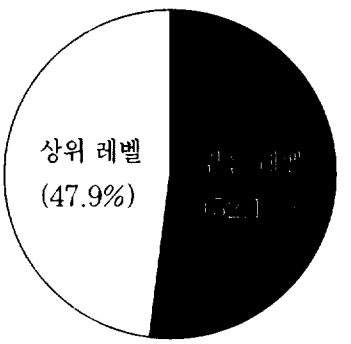
두가지의 신문에서 무작위로 추출한 266건의 신문기사에 대해서, 주제분석에 의해 얻어진 주제개념을 『신문기사 종합시소러스』의 디스크립터로 표현한 결과, 주제개념과 디스크립터간에 812의 표현관계가 얻어졌다. 단, 하나의 주제개념에 대해서 복수의 디스크립터를 합성하여 부여한 경우도 하나의 표현관계로 간주하였다.

812의 표현관계를 다시 보통명사와 고유명사로 나눈 결과, 보통명사는 587어로 전체 표현관계 중에서 72.3%를 차지하였고, 고유명사는 225어로 전체 중에서 27.7%를 차지하였다.

(1) 보통명사에 있어서 주제개념과 디스크립터와의 표현관계

보통명사에 있어서 주제개념과 디스크립터와의 표현관계는, 주제개념에 대해서 디스크립터가 표현할 수 있는 개념의 레벨으로 분류하였다. 즉, 디스크립터가 주제개념과 같은 레벨의 개념으로 표현하고 있는 경우와 디스크립터가 주제개념보다 넓은 상위레벨의 개념으로 표현하고 있는 경우이다. 후자의 경우, 디스크립터가 표현하고 있는 개념은 주제개념에 비해서 특정성이 낮다.

조사의 결과, 디스크립터와 주제개념이 같은 레벨의 표현관계는 306(52.1%), 디스크립터의 경우가 주제개념보다도 상위의 표현관계는 281(47.9%)였다(그림1 참고).

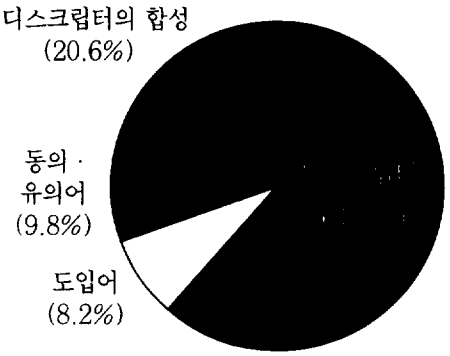


〈그림 1〉 같은 개념 레벨과 상위의 개념레벨의 비율

디스크립터가 주제개념과 같은 레벨의 개념을 표현하고 있는 경우(306)는 그 주제개념을 하나의 디스크립터로 표현하고 있는 경우

(243)와 복수의 디스크립터를 합성하여 표현하고 있는 경우(63)로 나뉘었다. 하나의 디스크립터로 표현하고 있는 경우는, 語形의 대응관계에 의해서, 셋으로 분류하였다. 첫째, 주제개념과 같은 어형의 디스크립터가 시소러스에 존재하는 경우이다. 예를들면, 주제개념인 〔거품경제〕에 대해서 같은 어형의 디스크립터가 존재하는 경우이다. 이 관계는 188가 있고, 같은 개념레벨의 표현관계 중에서 61.4%를 차지했다. 둘째, 주제개념과 같은 어형의 도입어가 있는 경우이다. 예를들면, 주제개념인 〔국제경제〕에 대해서 같은 어형의 도입어가 있고, 그곳에서 디스크립터인 〔세계경제〕로 지시되어 있는 경우이다. 이 관계는 25이며, 같은 개념레벨의 표현관계 중에서 8.2%이었다. 셋째, 디스크립터와 도입어와는 어형에서는 일치하지 않으나, 주제개념에 대해서 동의 · 유의어인 개념레벨의 디스크립터로 표현하는 경우이다. 예를들면, 주제개념인 〔교역〕에 대해서 어형이 같은 디스크립터와 도입어는 없으나, 같은 개념레벨인 〔무역〕이 디스크립터로서 있는 경우이다. 이 관계는 30이고, 같은 개념레벨의 관계 중에서 9.8%이었다.

하나의 디스크립터로는, 주제개념을 적절히 표현할 수 없기 때문에, 복수의 디스크립터를 합성하여 주제개념과 같은 레벨의 개념을 표현하는 경우도 있었다. 예를들면, 주제개념인 〔수입상품〕을 표현하는 디스크립터가 없기 때문에 〔수입〕과 〔상품〕의 두개의 디스크립터를 합성하여 주제개념을 표현하는 경우도 있다. 이 관계는 63이 있으며, 같은 개념레벨의 관계중에서 20.6%를 차지하였다(그림 2 참고).

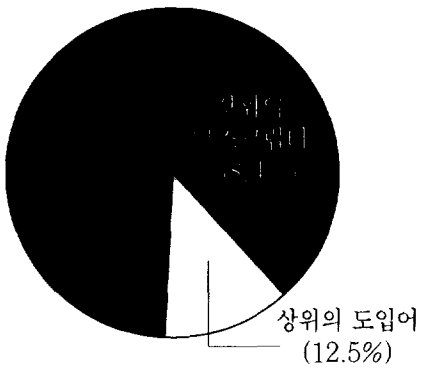


〈그림 2〉 같은 개념 레벨의 표현 비율

한편, 같은 레벨의 개념으로 표현하는 경우에 대해서, 디스크립터가 주제개념보다도 상위레벨의 개념을 표현하고 있는 경우(281)는, 표현이 許容範圍內의 경우와 許容範圍外의 경우로 나뉘어졌다. 표현이 許容範圍內의 경우(143)는 어형과의 대응관계에 의해서 두개로 나뉘어졌다. 첫째, 주제개념과 같은 어형의 도입어가 있고, 이 도입어가 보다 상위의 개념의 디스크립터를 지시하고 있는 경우이다. 예를들면, 주제개념인 〔노벨경제학상〕에 대해서 같은 어형의 도입어인 〔노벨경제학상〕이 있고, 이것이 상위개념의 디스크립터인 〔노벨상〕을 지시하고 있는 경우이다. 이 경우는, 주제개념에 비해서 디스크립터의 개념이 특정성이 낮다. 이 관계는 35가 있고, 상위의 개념레벨의 관계중에서 12.5%이었다. 둘째, 어형이 같은 도입어가 없어, 상위의 개념레벨의 디스크립터로 표현하는 경우이다. 예를들면, 주제개념인 〔경기회복책〕에 대해서, 상위의 개념레벨의 디스크립터인 〔경제정책〕을 부여하는 경우이다. 이 관계는 108이 있고, 상위의 개념레벨의 관계중에서 38.4%를 차지했다.

한편, 주제개념에 대해서 시소러스 중에서

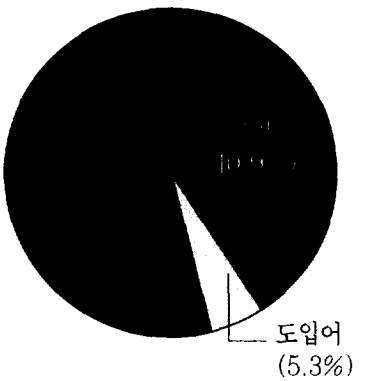
가장 가까운 상위개념인 디스크립터를 부여했지만, 디스크립터의 관련어가 많거나 주제개념과 너무 떨어져 있기 때문에, 주제개념에 비해서 디스크립터가 표현하는 개념이 너무 넓어서, 검색에 있어서 잡음이 생기기 쉽다고 생각되는 경우는 許容範圍外의 표현이라고 판단했다. 예를들면, 주제개념인 〔원산지〕에 대해서, 시소러스에서 가장 가까운 상위개념인 〔농산물〕이라는 디스크립터를 부여했지만, 이 경우는 주제개념에 비해서 디스크립터가 표현하는 개념이 너무 넓기 때문에 주제개념에 대해서 許容範圍外의 표현이라고 판단했다. 이 관계는 138이었고, 상위의 개념레벨의 관계 중에서 49.1%를 차지했다(그림 3 참고).



〈그림 3〉 상위 레벨의 표현

(2) 고유명사에 있어서 주제개념과 디스크립터와의 표현관계

고유명사에 있어서 주제개념과 디스크립터와의 표현관계는, 보통명사와 같이 개념의 레벨로 나뉘어지지 않기 때문에, 주제개념에 대응하는 디스크립터가 『신문기사 종합시소러스』에 있는가 없는가에 의해서 분류했다. 또한, 주제개념에 대응하는 디스크립터가 『신문기사 종합시소러스』에 있는 경우는 주제개념을 표현하는 말의 어형이 디스크립터의 어형과 일치하는 경우와 도입어의 어형과 일치하는 경우로 나뉘어졌다. 주제개념을 표현하는 말과 디스크립터와의 어형이 일치하는 경우는 92가 있고, 고유명사 전체 중에서 40.9%를 차지했다. 또한, 주제개념을 표현하는 말과 도입어의 어형이 일치하는 경우는 12가 있고, 고유명사 전체 중에서 5.3%이었다. 한편, 주제개념을 나타내는 말과 일치하는 디스크립터 혹은 도입어가 『신문기사종합시소러스』에 없는 경우는 121이며, 고유명사 전체 중에서 53.8%를 차지했다(그림 4 참고).



〈그림 4〉 고유명사에 있어서 주제개념과 디스크립터의 표현관계의 비율

3.2 신문용어 수집 리스트와의 대조 결과

신문기사로 부터 추출한 1149語 및 세종류의 신문 용어사전과 신문용어의 분류표로 된 네가지의 자료에서 셋이상에 수록되어 있는 293語를 합친 1442語의 수록용어를 『신문기사 종합시소러스』와 대조한 결과, 『신문기사 종합시소러스』와 일치하는 수집용어는 702語

〈표 1〉『신문기사 종합시소라스』와 수집용어 리스트와의 대조결과

용어의 수집원	추 출 어	일치의 수집어	누락의 수집어
신문기사	1149(79.7)	594(84.6)	555(75.0)
매일 동아 중앙 분류표	24(1.7)	20(2.8)	5(0.7)
매일 동아 중앙	212(14.7)	46(6.6)	164(22.2)
매일 동아 분류표	30(2.1)	23(3.3)	7(0.9)
매일 중앙 분류표	21(1.5)	16(2.3)	6(0.8)
동아 중앙 분류표	6(0.4)	3(0.4)	3(0.4)
총 계	1442(100)	702(100)	740(100)

( )안은 %

*매일:『경제신어사전』(매일경제신문사 1991)

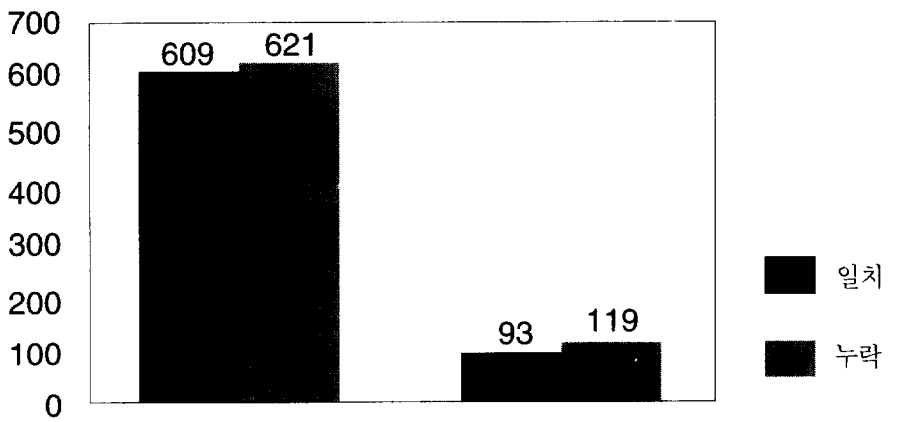
*동아:『현대시사용어사전』(동아일보사 1992)

*중앙:『현대지식정보사전』(중앙일보사 1992)

*분류표:『기사자료표준분류표』(한국언론연구원, 한국조사기사회 1991)

로 수집용어 전체 중에서 48.7%이며, 일치하지 않는 〔누락〕의 수집용어는 740語로, 수집용어 전체중에서 51.3%이었다. 단, 일치와 누락의 판단은 디스크립터 혹은 도입어의 어형과 수집용어의 어형이 일치하는가 하지 않는가에 의하여 하였고, 개념레벨과 동의 · 유의어에 대한 판단은 하지 않았다. 각각의 비율은 표1과 같다.

『신문기사 종합시소러스』와 수집한 신문용어 1442語를 대조한 결과, 얻어진 일치와 누락의 용어는 보통명사와 고유명사로 나뉘어졌다. 그 결과 보통명사 중(1230어)에서 일치가 609어(49.5%), 누락이 621어(50.5%)이었다. 한편, 고유명사 중(212어)에서는 일치가



〈그림 5〉 수집용어에 있어서 보통명사와 고유명사

〈표 2〉 보통명사에 있어서 일치와 누락

주 제 분 야		총 수	일 치	누 락
경 제	경 제	205(16.7)	60(9.9)	145(23.3)
	금 융	245(19.9)	93(15.3)	152(24.5)
	기 업	125(10.2)	57(9.4)	68(11.0)
	노 동	53(4.3)	30(4.9)	23(3.7)
	계	628(51.1)	240(39.4)	388(62.5)
산 업	산 업	75(6.1)	35(5.7)	40(6.4)
	농림수산	62(5.0)	34(5.6)	28(4.5)
	자 원	34(2.8)	30(4.9)	4(0.6)
	에너지	28(2.3)	21(3.4)	7(1.1)
	광공업	51(4.1)	40(6.6)	11(1.8)
	운수교통	29(2.4)	21(3.4)	8(1.3)
	토목건설	30(2.4)	21(3.4)	9(1.4)
	상 업	117(9.5)	53(8.7)	64(10.3)
	정보통신	46(3.7)	36(5.9)	10(1.6)
	계	472(38.4)	291(47.8)	181(29.1)
	그외의 분야	198(16.1)	146(24.0)	52(8.4)
	총 계	1230(100)	609 (100)	621(100)

일치에 중복어 68어 있음. ( )안은 %

93어(43.9%)이고, 누락이 119어(56.1%)이였다(그림 5 참고).

(1) 보통명사에 있어서 일치와 누락

보통명사에 있어서 일치와 누락의 용어를 『신문기사 종합시소러스』의 주제분야에서 의해서 분류하고, 그 비율을 조사하였다.(표2 참조)

일치와 누락중에서, 경제 · 산업의 각 분야의 비율을 보면, 일치의 609어중에서 경제분야는 240어(39.4%), 산업분야는 291어(47.8%)이었다. 단, 두 분야에 들어가는 중복어는 68어이었다. 경제 · 산업분야의 下位分野에 있어서의 비율을 조사한 결과, 경제분야의 하위분야에서는 금융분야의 비율이 가장 높은 15.3%였다. 또한, 산업분야의 하위분야에서는, 상업분야의 비율이 가장 높은 8.7%이었다. 한편, 누락의 621어중에서 경제분야는 388어(62.5%), 산업분야는 181어(29.1%)이었다. 하위분야에서는 경제분야의 하위분야인 금융분야의 비율이 가장 높은 24.5%이었다. 산업분야의 하위분야인 상업분야의 비율이 가장 높은 10.3%이었다.

(2) 고유명사에 있어서 일치와 누락

고유명사에 있어서 수집용어와 『신문기사 종합시소러스』의 대조에 있어서는, 고유명사를 국제분야와 국내분야로 나누고, 다시 고유명사

〈표 3〉 고유명사에 있어서 일치와 누락

고유명사의 종류		총 수	일 치	누 락
국 제	기관	21(9.9)	16(17.2)	5(4.2)
	협정	5(2.4)	2(2.2)	3(2.5)
	회의	5(2.4)	4(4.3)	1(0.8)
	경제협의체	9(4.2)	7(7.5)	2(1.7)
	국가	8(3.8)	8(8.6)	0(0.0)
	제도,규격	6(2.8)	2(2.2)	4(3.4)
	계	54(25.5)	39(41.9)	15(12.6)
국 내	단체	30(14.2)	11(11.8)	19(16.0)
	행정기관	19(9.0)	18(19.4)	1(0.8)
	정부출연기관	16(7.5)	5(5.4)	11(9.2)
	연구기관	14(6.6)	5(5.4)	9(7.6)
	금융기관	17(8.0)	7(7.5)	10(8.4)
	법,정책	22(10.4)	8(8.6)	14(11.8)
	기업	40(18.9)	0(0.0)	40(33.6)
	계	158(74.5)	54(58.1)	104(87.4)
총 계		212(100)	93(100)	119(100)

( )안은 %

의 종류별로 나누어 그 비율을 조사하였다.(表 3 참조)

결과, 『신문기사 종합시소러스』와 일치하는 수집용어의 93어중에서, 국제분야는 39어(41.9%)를, 국내분야는 54어(58.1%)이었다. 국제분야 중에서는 국제기관명이 16어(17.2%)로 가장 높은 비율이었다. 국내분야 중에서는, 행정기관명이 18어(19.4%)로 가장 높은 비율을 차지하였다.

한편, 누락의 수집용어의 119어 중에서, 국제분야는 15어(12.6%), 국내분야는 104어(87.4%)이었다. 국제분야 중에서는, 국제기관명이 5어(4.2%)로 가장 높은 비율이었다. 국내분야 중에서는, 회사명이 40어(33.6%)로 가장 높은 비율이었다.

4. 고 찰

4.1 주제개념에 대한 표현력

(1)보통명사에 있어서 같은 개념레벨의 표현

『신문기사 종합시소러스』의 디스크립터가 신문기사의 주제개념을 어느 정도 적절히 표현할 수 있는가 라는 표현력의 평가를 위해, 색인작업에 의해서 얻어진 표현관계에 대해서 보통명사와 고유명사로 나누어 분석을 하였다.

보통명사에 있어서 587의 표현관계를 개념

의 레벨에 의해서 나눈 경우, 같은 개념의 레벨의 표현관계에 있어서, 주제개념을 표현하는 말과 도입어와 어형이 일치하는 관계는, 같은 개념레벨의 관계중에서 가장 낮은 비율(8.2%)이였다. 이번 색인작업의 결과 나타난 도입어 중에서는, 신문기사에 디스크립터보다 출현한 빈도가 높은 도입어도 보였다. 그 때문에, 도입어와 디스크립터간의 우선관계를 재조정하는 것은 신문 시소러스의 관리에 있어서 고려해야 할 중요한 문제일 것이다.

어형이 일치하는 디스크립터와 도입어가 없기 때문에, 주제개념과 동의 · 유의어인 디스크립터로 표현하는 관계는 도입어와 어형이 일치하는 경우를 조금 상회(9.8%)하였다. 디스크립터와 도입어로서 통제되지 않은 동의 · 유의어가 많으면 용어가 분산되기 때문에 색인작업에 있어서 일관성을 갖기 어렵고, 검색시에도 누락되어 버리는 문제가 생긴다. 따라서, 동의 · 유의어 중에서 출현빈도가 높은 것은 시소러스에 도입어나 디스크립터로 선택하여 통제를 할 필요가 있을 것이다.

한편, 『신문기사 종합시소러스』에서는 주제개념을 디스크립터가 같은 레벨에서 표현하는 경우, 하나의 디스크립터를 표현하는 경우가 복수의 디스크립터를 합성하여 표현하는 경우의 비율을 크게 상회하여, 디스크립터의 事後結合에 의한 주제개념의 표현력은 그다지 높다고 할 수 없다.

(2) 보통명사에 있어서 상위의 개념레벨의 표현

신문기사의 주제개념을 같은 개념레벨로 표현하는 디스크립터가 없는 경우는 상위개념의 디스크립터를 부여하였다. 상위개념의 디스크립터로 표현하는 경우는 주제개념에 대해서 시소러스 중 가장 가까운 상위의 디스크립터를 부여하였으나, 그 개념이 너무 넓기 때문에 주제개념을 적절하게 표현하지 못하는 경우도 포함되어 있었다. 그 때문에, 부여한 디스크립터가 許容範圍內의 표현인가 許容範圍外의 표현인가를 판단하여 분류하였다. 단, 허용범위에 대한 판단은 主觀이 개입되기 때문에 그 비율은 그다지 엄밀하다고 할 수는 없다. 그 결과, 상위의 개념레벨의 281의 관계중에서 허용범위내의 표현은 50.9%(143)이었고, 허용범위외의 표현은 49.1%(138)이었다. 이 중에서, 허용범위외의 표현은 『신문기사 종합시소라스』의 평가에 있어 가장 문제가 되는 표현관계로서, 보다 자세히 분석할 필요가 있다.

먼저, 허용범위외의 표현관계에 있어서, 주제개념을 나타내는 말을 單語와 복수의 構成要素로 되어 있는 複合語로 나누면, 허용범위외의 표현관계 중에서 단어는 18.1%(25), 복합어는 81.9%(113)로서, 복합어가 훨씬 높은 비율을 차지하였다. 이와같이, 허용범위외의 표현관계에 있어서 주제개념을 표현하는 말 중에서 8할 이상을 차지한 복합어에 대해서, JIS X0901-1991의 복합어 분리기준에 의거하여, 構成槪念으로 분리하는 편이 좋은 경우와 분리하지 않은 편이 좋은 경우로 나누었다.

그 결과, 분리하는 편이 좋다고 판단되는 주제개념을 나타내는 말은 복합어로 표현된 관계(113) 중에서 52.2%(59)이었으며, 분리하지 않는 편이 좋다고 판단되는 말은 47.8%(54)이었다. 이와같이 허용범위외의 표현관계의 경우에 주제개념을 나타내는 말이 單語보다 複合語가 훨씬 많고, 복합어 중에서는 분리하는 편

이 좋은 경우가 분리하지 않는 편이 좋은 경우의 비율을 상회하였다.

여기서, 허용범위외로 표현된 주제개념을, 單語및 複合語로서 분리하지 않는 편이 좋은 경우(주제개념을 분리하지 않는 경우로 부름)와, 복합어로서 분리하는 편이 좋은 경우(구성개념으로 분리하는 편이 좋은 경우로 부름)로 나누어 검토하기로 한다.

먼저, 주제개념을 분리하지 않는 경우는 허용범위외의 표현관계 전체(138)중에서 57.2%(79)이었다. 이 표현관계를 다시 셋으로 나누어보면, 첫째, 주제개념에 대해서 부여된 디스크립터가 너무 상위의 개념으로서 주제개념과 떨어져 있기 때문에, 보다 특정성이 높은 개념을 나타내는 디스크립터가 필요하다고 판단되는 경우로, 허용범위외의 표현관계 중에서 24.6%(34)의 비율이었다. 둘째, 주제개념의 구성요소의 일부가 디스크립터로 있고, 주제개념보다도 상위의 개념을 그 디스크립터로 어느 정도는 표현할 수 있지만, 주제개념을 나타내는 말 자체가 신문기사에 잘 나타나고 일반적으로 친숙해져 있기 때문에, 그 디스크립터를 부여하는 것 보다는 주제개념을 표현하는 디스크립터를 시소러스에 도입하는 편이 좋다고 판단되는 경우이다. 이 경우는 허용범위외의 표현관계 중에서 28.3%(39)였다. 셋째, 적절한 디스크립터가 없는 전문용어와 新語의 경우로서, 허용범위외의 표현관계 중에서 가장 적은 4.3%(6)이었다.

이상과 같이, 주제개념을 분리하지 않는 경우는 허용범위외의 표현관계 중에서 높은 비율(57.2%)를 차지하고 있다. 이들의 허용범위외의 개념을 보다 주제개념에 가깝게 표현할 수 있도록 하기 위해서는, 이번 색인작업에서 부여한 디스크립터 보다 특정성이 높은 하위레벨의 디스크립터가 필요하다고 할 수 있다.

다음, 구성개념으로 분리하는 편이 좋은 경우는, 허용범위외의 표현관계 전체(138) 중에서 42.8%(59)이었다. 이번의 색인작업에서 분리할 수 있는 복합어를 분석한 결과, 구성개념으로 분리하는 편이 좋은 경우(59)중에서, 95%(56)가 구성개념의 일부가 이미 존재하는 디스크립터로 표현할 수 있는 경우였고, 표현할 수 없었던 구성개념의 부분을 디스크립터로 도입하면, 보다 표현력을 높일수 있다는 것이 보여졌다. 이것으로 부터, 허용범위외의 주제개념에 대응하는 디스크립터를 시소러스에 도입할 때는 분리할 수 있는 복합어는 각 구성개념에 대응하는 디스크립터로 넣는 것이 좋을 것이다.

한편, 이들의 허용범위외의 표현관계가 포함되어 있는 주제분야로서, 본 연구의 대상분야인 경제 · 산업분야에서는 경제분야가 73으로 산업분야의 38의 약 2배에 달하였다. 하위분야에서는 경제분야의 금융과 산업분야의 상업분야가 많이 나타났다. 그러므로 이들의 분야의 주제개념을 표현하는 디스크립터의 도입에 대해서 특히 검토할 필요가 있을 것이다.

(3) 고유명사

색인작업의 결과로 부터 얻어진 812의 표현관계 중에서 고유명사는 27.7%(225)를 차지하여, 고유명사가 신문기사의 중요한 개념으로서 높은 비율을 나타내고 있다는 것이 보여졌다.

『신문기사 종합시소러스』에서는 국제분야의 고유명사에 대해서는 상당히 망라적으로 다루

고 있으나, 국내분야에 대해서는 전반적으로 누락이 많았다. 그 중에서 신문기사에서 특히 많이 나타나는 회사명은 검색의 실마리가 되는 중요한 키워드로 될 수 있는 주제어로서 신문기사에 빈번히 출현함에도 불구하고, 『신문기사 종합시소러스』에서는 다루고 있지 않았다.

또한 『신문기사 종합시소러스』의 고유명사는 대부분 正式名을 도입어로 넣고 略語를 디스크립터로 하는 경우가 많았으나, 약어가 도입어로서 나타난 경우도 보였다. 고유명사는 정식명과 약어 양쪽이 신문기사에서 잘 나타나는 경우가 많기 때문에 정식명과 약어 중에서 어느쪽을 디스크립터로 할 것인라는 고유명사의 수록방법은 신문 시소러스에 있어서 고려해야 할 문제라 할 수 있다

4.2 주제분야에 대한 포괄성

(1) 보통명사

『신문기사 종합시소러스』에 할당되어 있는 색인어의 비율은 산업분야가 경제분야보다 2배정도가 많고, 각 하위분야에 할당되어 있는 용어도 분야별 차가 별로 없었던 것에 비해서, 수집용어에서는 경제분야가 산업분야보다 많고, 각 하위분야에 있어서도 금융분야와 상업분야의 용어수가 다른 하위분야보다 대폭 많아, 수집용어와 『신문기사 종합시소러스』의 색인어간에 주제분야별 배분에 相異가 보여졌다.

또한, 『신문기사 종합시소러스』를 수집용어와 대조한 결과, 누락된 용어 중에는 경제분야의 용어의 비율이 산업분야의 용어의 누락의 비율을 대폭 상회하였다. 한편, 하위분야에 있어서는 금융분야와 상업분야의 비율이 다른 하위분야의 비율보다도 높았다. 이와같이, 『신문기사 종합시소러스』에서는 신문기사에서 잘 쓰여지는 주제분야의 용어에 누락이 많은 것이 보여졌다. 이것으로 신문에서 자주 출현하는 주제분야의 용어를 보다 많이 수록하는 것은 신문 시소러스가 주제분야를 포괄하는데 있어서 중요한 문제라고 할 수 있다.

(2) 고유명사

고유명사를 어느 정도 넓게 포괄할 수 있는가를 살펴보기 위해서 수집한 고유명사를 국제분야와 국내분야로 나누고, 고유명사의 종류별로 나누었다. 고유명사의 수집용어 225어 중에서는 국내분야의 비율이 국제분야의 비율보다 높았으나, 『신문기사 종합시소러스』와 대조한 결과에서는, 누락의 용어 중에서 국내분야의 비율이 국제분야의 비율보다 대폭 높았다.

이것으로 『신문기사 종합시소러스』가 국제분야의 고유명사에 대해서는 포괄적으로 수록하고 있다고 할 수 있으나, 국내분야의 고유명사에 대해서는 포괄적이지 못하다는 문제가 보여졌다.

5. 결 론

본 연구는 신문 시소러스의 평가에 관한 연구의 일환으로서, 한국에서 처음 만들어진 신문 시소러스인 『신문기사 종합시소러스』의 경제 · 산업분야를 대상으로 색인작업과 신문용어의 수집 리스트와의 대조라는 두가지의 방법으로 평가및 분석을 하였다. 그 결과, 신문 시소라스를 작성하고 관리하는데 있어서 다음 다

섯가지의 점이 특히 중요하다는 것이 밝혀졌다.

첫째, 색인어의 特定性의 문제이다. 색인작업의 결과, 전체 표현관계 중에서 주제개념을 같은 개념레벨의 디스크립터로 표현하지 못하고 상위의 개념레벨의 디스크립터로 표현하는 경우의 비율이 높고, 상위의 개념레벨의 디스크립터가 주제개념을 허용범위외에서 표현하는 경우의 비율이 높았다. 더구나, 허용범위외에서 표현되는 주제개념에는 구성개념으로 분리될 수 있는 복합어보다도 단어및 분리할 수 없는 복합어가 많았다. 이들의 주제개념을 적절히 표현하기 위해서, 기존의 디스크립터보다도 특정성이 높은 디스크립터를 시소러스에 도입할 필요가 있다고 판단된다. 색인어의 특정성은, 신문이 일반지인가 전문지인가에 따라서 혹은 취급하는 주제분야와 정보량에 의해서 레벨의 相異가 있는 것이 당연하나, 신문 시소러스에 있어서, 색인어가 주제개념을 적절히 표현할 수 있는 특정성을 갖는 것은 매우 중요한 문제라고 할 수 있다.

둘째, 複合語의 문제이다. JIS X0901에서는 원칙적으로 복합어는 구성개념으로 분리하는 것을 권하고 있는데, 이번 색인작업의 결과에서도 허용범위외에서 표현된 주제개념 중에서 분리할 수 있는 복합어를 분석한 결과, 대부분의 복합어는 구성개념의 일부가 이미 존재하는 디스크립터로 표현할 수 있는 경우였고, 표현할 수 없었던 구성개념의 부분을 디스크립터로 도입하면 보다 표현력을 높일 수 있다는 것이 보여졌다. 이것으로 부터 신문 시소러스에 있어서, 광범위한 신문주제를 표현하는데 있어서 표현력을 보다 높이기 위해서는, 복수의 디스크립터를 합성하여 표현할 수 있도록 복합어를 분리할 필요가 있다고 할 수 있다.

셋째, 導入語의 문제이다. 색인작업의 결과, 전체 표현관계 중에서, 도입어와 어형이 일치하는 표현관계의 비율은 가장 낮고, 도입어와 디스크립터간의 우선관계에 문제가 있는 것이 보여졌다. 신문 시소러스와 같이 방대한 용어를 다루는 시소러스에 있어서, 도입어는 디스크립터로서 선정되지 않은 동의 · 유의어와 보다 특정성이 높은 용어를 디스크립터에 안내하는 것에 의해 일관된 색인작업을 하고 검색의 재현성을 높이기 위해 중요한 문제라고 할 수 있다.

넷째, 固有名詞의 문제이다. 고유명사는 신문기사의 주제요소로서 신문 정보의 검색에서는 중요한 액서스 포인트가 되지만, 시소러스에 수록될 수 있는 수는 제한되며, 고유명사 중에서 많이 포함되어 있는 略語와 新語에 대해 색인작업을 어떻게 할 것인가라는 문제가 있으므로, 신문 시소러스의 작성과 관리에 있어서 고유명사의 취급은 중요한 문제라고 할 수 있다.

다섯째, 주제분야간의 用語配分의 문제이다. 『신문기사 종합시소러스』를 수집용어 리스트와 대조한 결과, 누락된 용어 중에서 신문기사에서 잘 나타나는 주제분야의 용어의 비율이 높다는 것으로 부터, 주제분야간의 용어의 배분에 문제가 보여졌다. 신문이 전문지인가 일반지인가에 따라서 주제분야에 대한 분류체계와 할당된 주제분야의 용어의 수는 다르므로, 신문기사의 중심영역의 주제분야의 개념을 보다 포괄적으로 표현할 수 있기 위해, 주제분야간의 용어의 배분은 신문 시소러스의 작성에 있어서 고려해야 할 중요한 문제라고 할 수 있다.

이상, 본 연구에서는 일반지를 대상으로 한

신문 시소러스인 『신문기사 종합시소러스』를 평가했으나, 이외에도 전문지를 대상으로 한 신문 시소러스와 다른 언어의 신문 시소러스에 대한 평가를 통해서, 신문 시소러스의 특성과 문제에 관한 연구가 더욱 축적되어 가야 할 필요가 있다고 생각된다.

참고문헌

松尾光, 神尾達夫. 日本における新聞記事データベースの現狀と今後の動向. 情報管理. vol.35, No. 10,p.872(1993)

阿部哲也. 中日新聞記事データベース : ACE-CHUNICHIの構築. 情報管理. Vol. 28, No. 2, p.117(1985)

Lancaster, F. W. "索引·探索作業における統制語彙の役割" 情報檢索の言語. 松村多美子譯. 東京, 日本 ドクメンテーション協會, 1976, p. 393-405

神尾達夫. 新聞記事データベースにおけるキーワード自動抽出. 情報管理. Vol. 32, No. 4, p. 284(1989)

菊池敏典, 染谷浩司. 分類·索引·シソーラスの知識. 情報管理. Vol. 35, No. 3, p. 225-237(1992)

廣木守雄, 阿部哲也. ニュース·シソーラス—ニュースをとらえる言葉·情報管理. Vol. 34, No. 1, p. 72-75 (1991)

羽原隆司. 日經産業ファイル—その特徵と活用法. 情報管理. Vol. 35, No. 8, p.678(1992)

神尾達夫. 新聞記事データベースにおけるキーワード自動抽出. 情報管理. Vol. 32, No. 4, p. 286-287(1989)

日本工業標準調査會. シソーラスの構築及びその作業方法 : JIS X 0901. 東京, 日本規格協會, 1991, 1p.

Lancaster, F. W. Information Retrie-val Systems : Characteristics, Testing and Evaluation. 2nd ed. John Wiley, 1979, 133p.

Kristensen, J. ; Jarvelin, K. The effectiveness of a searching thesaurus in free-text searching in a full-text database. International Classificaion. vol. 17, No.2,p.77-84(1990)

Sager, J. C. ; Somers, H. L. ; McNaught, J. Thesaurus integra-tion in the social sciences, pt.1 : Comparison of thesauri. Interna-tional Classification. Vol.8, No.3, p. 133-138(1981)

Sager, J. C. ; Somer, H. L ; Mcnaught, J. Thesaurus inter-gration in the social sciences, pt. 2 : Stages towards intergration. International Classification. Vol. 9, No. 1, p. 19-26(1982)