

한·일 인터넷 정보유통을 위한 이용자 친화적 다국어 시소러스 모형 개발에 관한 연구*

Modeling a User-Friendly Korean-Japanese Thesaurus for Internet Circumstances

한 승 희(Seung-Hee Han)*

초 록

본 연구의 목적은 이용자 친화적인 한·일 시소러스 구축을 위한 모형을 개발하는 것이다. 이를 위해 선행연구 분석과 한국과 일본의 시소러스 구축 경험자 인터뷰를 통해 기존의 다국어 시소러스 구축방법에서 제기될 수 있는 문제점을 분석하고 현재의 정보환경에서 언어적 혹은 문화적 차이에 따른 이용자들의 어휘사용의 차이를 확인하기 위해 한국과 일본의 이용자 태그를 분석하였다. 분석결과, 인터넷 정보유통 환경에서 적합한 시소러스 모형인 탐색 시소러스 모형을 확인하고, 이를 적용하여 이용자 친화적인 형태의 한·일 시소러스 모형을 제안하였다.

ABSTRACT

The purpose of this study is modeling a user-friendly Korean-Japanese thesaurus for internet circumstances. The methodology had two steps: first, the problems were analyzed when multilingual thesaurus is constructed, with review of various multilingual thesauri and interview with multilingual thesaurus constructor in Korea and Japan. Second, collaborative tags were analyzed to differentiate cultural/linguistic background from word usage pattern. As a result, searching thesaurus model was suggested and applied to develop a user-friendly Korean-Japanese thesaurus.

키워드: 다국어 시소러스, 한·일 시소러스, 다국어 탐색 시소러스 모형, 이용자 친화적 시소러스
Multilingual Thesaurus, Korean-Japanese Thesaurus, Multilingual Searching Thesaurus
Model, User-Friendly Thesaurus

* 이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2004-037-H005).

** 서울여자대학교 문헌정보학과 전임강사(hanshee@swu.ac.kr)

논문접수일자 : 2007년 11월 21일 논문심사일자 : 2007년 11월 24일 게재확정일자 : 2007년 12월 4일

1. 서론

1.1 연구의 목적 및 필요성

전통적인 인쇄매체의 정보환경에서부터 인간은 지식에 대해 효과적으로 접근하기 위해 다양한 지식구조 표현기법을 개발해 왔다. 그 결과로 정보 이용자는 전체 학문분야나 특정 주제분야에 대해 구조적으로 접근하기 위해 분류표, 시소러스, 주제명 표목 등과 같은 전역적 수준의 지식구조(global-level knowledge structure)와 특정 저작 내에서 발생한 지식에 구조적으로 접근하기 위해 목차와 권말색인 등과 같은 지역적 수준의 지식구조(local-level knowledge structure)를 이용할 수 있게 되었다(한승희 2004).

이러한 지식구조 표현에 대한 관심은 다국어 환경에서도 예외가 아니다. 다국어 정보환경에서의 지식에 대해서도 구조적 표현과 접근에 대해 연구할 필요가 있는데, 그 이유는 다음과 같다.

첫째, 전 세계적으로 진행되고 있는 국제화 추세와 초고속 정보통신 기술의 발전으로 인한 인터넷의 대중화로 인해 정보 이용자들은 다양한 언어로 된 정보에 쉽게 접근할 수 있게 되어, 이용자는 자신의 정보 요구를 충족시키기 위해 자국어로 된 정보 이외에 외국어로 된 정보를 필요로 하게 되었다. 외국어로 된 정보에 대한 요구는 이미 오래전부터 국제적 수준의 학술정보 유통과정에서부터 발생된 것으로, 학술연구정보의 효과적인 공유를 위해 전산학 분야와 문헌정보학 분야 연구자들을 중심으로 하여 다국어 시소러스(multilingual thesaurus), 교차언어 정보검색(cross language information retrieval

CLIR), 자동번역 시스템(automatic translation system) 등을 주제로 한 많은 연구들이 수행되어 왔다.

둘째, 최근 우리나라에서는 일본문화 개방으로 인해 일본어로 된 정보에 대한 이용자의 관심이 높아지게 되었고, 일본에서는 한류 열풍과 함께 한국어로 된 정보에 대한 이용자의 관심이 높아지게 되었다. 이에 따라 한국과 일본 간의 정보유통체계를 효과적으로 확립하기 위한 기초적인 단계로 언어적인 문제를 해결할 필요가 있다.

다국어 시소러스란 둘 이상의 언어로 구성된 시소러스로서 용어의 상호관계 및 각 언어의 등가어를 배열하고 있는 시소러스를 의미한다(ISO 5964). 이것은 여러 개의 단일어 시소러스를 함께 모아놓은 것 이상의 것으로, 반드시 개념간의 의미구조를 취해야만 하며, 원문언어(source language)에 대해 목적언어(target language)로 번역 대응어 디스크립터만 작성하고 의미구조를 제공하지 않는 시소러스는 진정한 의미의 다국어 시소러스라고 할 수 없다(Hudon 1997).

다국어 시소러스는 원문언어는 물론 목적언어의 범위에 포함되는 대규모의 정보를 효율적으로 검색할 수 있도록 도와주는 탐색보조도구(search aids)로 사용될 수 있다. 뿐만 아니라, 다국어 번역에 있어 번역자의 주관적이고 임의적인 판단에 의존하던 번역어 선택의 문제를 상당히 해소하고, 번역용어의 표준화를 위한 기초자원으로 활용될 수 있으며, 자동번역 시 전문용어 전자사전으로 기능할 수 있다.

다국어 시소러스의 필요성은 대부분 정보가 인터넷을 통해서 제공되면서 더욱 커졌다고 할

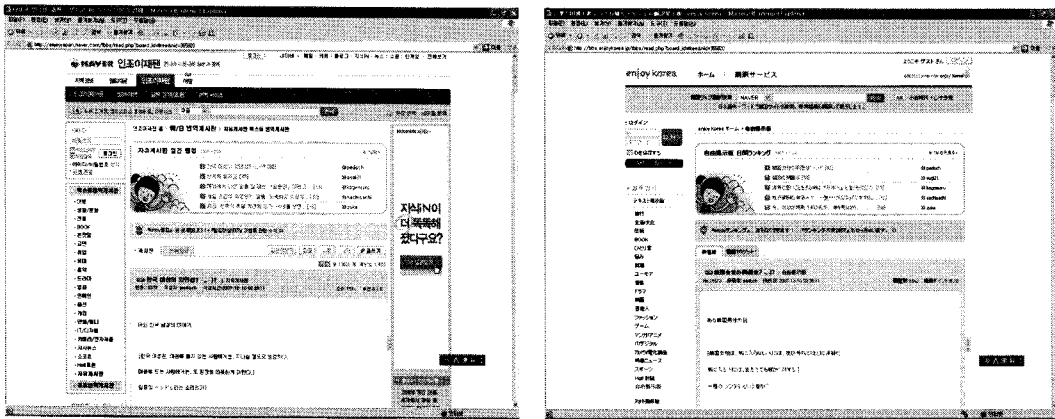
수 있다. 기존 도서관의 온라인 전자저널을 근간으로 한 전자도서관 서비스와 학술정보포털 서비스 등이 모두 인터넷을 통해 제공되고 있다. 이외에도 다수의 정보 이용자들은 인터넷 검색엔진을 이용하여 학술적이거나 비학술적인 정보를 얻고 있다. 이러한 현상을 반영하듯, 현재 인터넷 포털 서비스 업체인 '네이버(http://www.naver.com)'에서는 한국에 '인조이재팬(http://enjoyjapan.naver.com/)', 그리고 일본에 'enjoykorea(http://bbs.enjoykorea.jp/)'라는 서비스를 제공하고 있는데, 이 서비스에서는 양국 언어로의 번역 서비스를 기반으로 한국과 일본의 다양한 정보를 양국 네티즌이 실시간으로 공유할 수 있게 하고 있다.

한편, 미국정보화표준기구(National Information Standard Organization)에서는 1993년에 발표된 ANSI/NISO Z39.19: Guidelines for the Construction, Format, and Management of Monolingual Thesauri를 2005년에 개정한 바 있다. 미국 내에서 널리 활용되고 있는 이 지침에 대해 개정을 시도한 이유는 바로 정보

이용 환경의 변화에서 찾아볼 수 있다.

과거의 정보 이용자들은 자신이 원하는 정보를 직접 탐색하기 어려웠기 때문에 사서나 색인전문가와 같은 정보전문가가 시소러스를 구축하고 이를 정보검색에 이용하였다. 그러나 온라인 정보검색이 도입되고, 접근 가능한 웹사이트나 데이터베이스가 증가하면서 최종 이용자가 시소러스와 같은 통제 어휘집을 직접 이용하여 원하는 정보를 검색할 수 있게 되었다. 이에 따라 정보 시스템 개발자들은 최종 이용자(end-user)의 정보검색을 지원할 수 있는 시소러스의 가치를 중요하게 인식하기 시작하였다(한승희 2006).

다국어 시소러스의 구축에 관한 국제 표준인 ISO 5964: 1985(E) Documentation - Guidelines for the Establishment and Development of Multilingual Thesauri는 다국어 시소러스 구축에 관한 한 전세계적인 표준으로 받아들여지고 있는 실정이나, 이 지침 역시 발표된 지 20여 년이 흘러 현재의 정보이용환경을 제대로 반영하고 있는지에 대해 의문을 제기할 수 있다.



(그림 1) 동일한 내용을 양국 언어로 제공하고 있는 '인조이재팬'과 'enjoykorea'

이러한 상황을 용약해보면, 현재의 정보환경에서는 다국어 시소러스의 중요성은 부각되어 있으나 20여 년 전의 지침을 따라 다국어 시소러스를 작성하기에는 무리가 있다고 할 수 있다.

시소러스를 구축할 때 고려해야 할 두 개의 큰 축이라 할 수 있는 어휘의 '통제'와 이용자의 '이용' 중 무엇을 중심으로 시소러스를 구축하는가에 따라 시소러스의 성격은 크게 달라질 수 있다(한승희 2006). 본 연구에서는 기존의 시소러스 연구 양상이라 할 수 있는 용어의 '통제' 뿐만 아니라 이용자의 '이용'까지 고려한 이용자 친화적인 다국어 시소러스를 구축하는 것을 목적으로 한다. 구체적으로는, 한국과 일본 간의 문화개방을 기점으로 두 국가간의 효율적인 정보 교류의 필요성이 대두되면서, 언어의 장벽에서 오는 정보이용 문제들을 해소할 수 있는 정보유통 환경을 마련하기 위한 한 방안으로 쌍방향 정보유통이 가능한 이용자 친화적인 한·일 다국어 시소러스의 모형을 제안하고자 한다.

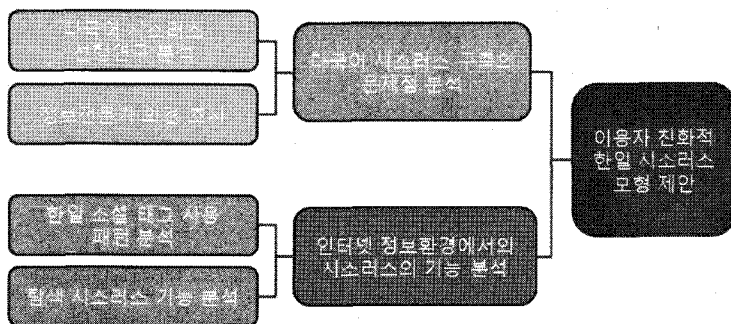
1.2 연구의 내용 및 방법

본 연구의 방법은 다음 그림 2와 같이 요약될 수 있다.

이용자 친화적 한·일 시소러스 모형을 제안하기 위해 수행된 연구의 내용은 크게 두 가지로 나누어볼 수 있다.

첫째, 기구축된 다국어 시소러스와 다국어 시소러스 구축 표준을 포함한 관련 선행연구를 분석하고, 시소러스의 구축 및 이용과 관련된 정보전문가의 의견을 인터뷰를 통해 조사하여 다국어 시소러스 구축의 문제점을 분석한다.

둘째, 분석된 문제점을 해결하기 위한 방안으로 인터넷 정보환경에서의 시소러스의 기능을 분석한다. 인터넷 환경에서 시소러스를 적용함에 있어 가장 적절하다고 판단된 탐색 시소러스의 기능을 분석하고, 또한 정보이용자의 관점에서 각국의 인터넷 정보이용, 특히 정보를 표현하는 어휘의 패턴을 분석하기 위해 한국과 일본의 소셜 북마킹 서비스에 사용된 이용자 태그를 분석하여 이용자 친화적인 탐색 시소러스에의 적용방안을 모색한다.



(그림 2) 이용자 친화적 한·일 시소러스 모형 개발 과정

2. 다국어 시소러스의 구축과 문제점

2.1 다국어 시소러스 구축현황

다국어 시소러스에 대한 본격적인 연구는 1990년대 중반 유럽지역을 중심으로 시작되었다. 대표적인 연구로는 1996년 시작된 *EuroWordNet* 프로젝트를 들 수 있다. 이 프로젝트는 *WordNet 1.5*에 덴마크어, 독일어, 이탈리아어, 스페인어, 불어, 체코어, 에스토니아어 등 유럽 7개 언어를 대상으로 단어간의 의미관계를 추가하여 대규모 다국어 시소러스를 구축하였다. 이외에도 현재 다국어 시소러스가 구축되어 실제로 검색시스템에 이용되고 있는 대표적인 시소러스로는 농학분야의 *AgroVoc*, 환경분야의 *GEMET*, EC의 활동과 관련된 분야를 다루는 *EuroVoc* 등을 들 수 있다(최석두, 조혜민 2002).

19세기 말부터 다국어 처리에 관심을 보이기 시작한 일본은 어학을 포함한 여러 분야에 시소러스가 구축되어 있는데, 전문통역자 단체가 만든 EDP 등의 다국어 시소러스와 국가적으로 추진하고 있는 다국어 전자사전인 『概念辭書(EDR 1995)』과 함께 『分類語彙表(1996)』 등이 돋보인다.

우리나라에서는 1980년대 초부터 교육·농업·노동·과학·언론 등의 분야에서 단일어 시소러스가 구축되어 왔으며, 이외에도 전문용어사전의 구축 노력이 여러 분야에서 수행되었다. 그러나 다국어 정보검색을 위한 기반기술 연구나 자원 개발은 위에서 언급한 국가들에 비해 매우 뒤떨어져 있는 상황이다.

최근 들어 '21세기 세종계획' 프로젝트에서

전문용어언어공학연구센터(KORTERM)를 운영하여 물리·화학·경제 분야의 한·일·영의 3개국 대응어 구축 및 표준화 작업이 이루어져 그 성과의 일부가 공개되었으며, 정영미 등(2001)은 과학기술 분류표, 시소러스, 용어사전 등을 한국어, 영어, 일본어의 3개 언어로 통합 구축하여 활용할 수 있도록 통합개념체계의 프로토타입을 구축하였다. 이외에도 차재은(2001)과 이형대(2001)는 한국학 분야를 대상으로 어휘 데이터베이스의 구축 및 활용방안과 다국어 시소러스 구축 방법론에 대해 각각 연구하였으며, 한국전자통신연구원(ETRI)에서 한·중 교차언어 검색 및 자동번역용 대역 전자사전을 개발하였다.

2.2 선행연구의 분석

국내에서 다국어 시소러스의 구축과 관련한 연구는 주로 양언어의 어휘적 통합이 주를 이루었다(한유석 2000, 고은숙 2007). 특히 한유석(2000)은 한일 양언어 시소러스 구축에 대해 여러 연구를 수행하였는데, 한국의 『연세한국어사전』과 일본의 『分類語彙表』를 대상으로 양언어의 어휘를 비교·분석하여 시소러스를 구축하고자 하였다. 이외에도 남영준(2001)은 지식관리시스템을 위한 의미형 한영 시소러스 구축에 관한 연구에서 용어 수집을 위해 기존의 문헌형태의 정보원과 웹과 같은 외부문서를 동시에 이용하였다.

외국의 경우 교차언어 정보검색과 자연어 처리를 위한 다국어 시소러스와 병렬 말뭉치(parallel corpus)의 자동 생성과 관련된 부분에 많은 연구가 있었다. Yang and Li(2003)의 연구에서는

영·중 병렬 말뭉치를 자동 생성하였고, Chen, Chau, Yeh(2003)는 병렬 텍스트 식별 시스템을 이용하여 영어와 중국어 웹 사이트를 대상으로 웹에서의 병렬 텍스트 탐색에 관한 연구를 수행하였다. 또한 Maeda, et al.(2000)의 연구에서는 검색엔진에서의 일·영 교차언어 정보검색을 위해 용어사전을 기반으로 한 질의번역을 통해 질의어의 의미모호성을 해소함으로써 정보검색의 성능을 높이고자 하였다.

선행연구의 분석결과, 기존의 다국어 시소러스 구축과 관련된 연구들은 기존의 다국어 시소러스 지침을 따르는 이외에도 전산학적 방법을 취하거나 어휘수집 방식에 변화를 주는 등 다양한 방법들을 적용하는 시도를 해온 것으로 알 수 있었다. 특히 대부분의 연구에서 중점을 두고 고려했던 점은 다국어 시소러스를 구축하는 데 있어 많은 논의가 되어왔던 원문언어와 목적언어 간의 언어적 문제를 해결하는 방법에 대한 것이었다. 일반적으로, 원문언어와 목적언어 간의 언어적 문제는 두 가지 관점에서 설명할 수 있다. 하나는 두 용어간의 등가어(equivalence) 처리에서 오는 문제이고, 다른 하나는 용어 간의 관계를 설명하는 것과 관련된 것이다.

ISO 5964에서는 다국어 시소러스를 구축할 때 발생하는 언어적·의미적 문제 중 가장 중요한 것이 대상 언어들의 모든 개념을 동일화하는 작업이며, 일반적으로 개념의 동일화와 관련하여 일어나는 다섯 가지 현상을 완전일치(exact equivalence), 부분일치(partial equivalence), 불완전일치(inexact equivalence), 일대다일치(single-to-multiple equivalence), 불일치(non-equivalence)라고 하였다(ISO 5964).

이 중에서 다국어 시소러스를 구축할 때 완전일치를 제외하고 일어날 수 있는 다른 현상들을 해결하기 위한 방법들을 여러 연구에서 제시하였다(Hudon 1997, 한유석 2000, 남영준, 안동언 2000, 남영준 2001, 이형대 2001).

용어간의 관계표현에 대한 문제점은 다국어 시소러스뿐만 아니라 단일어 시소러스 구축에 있어서도 나타난다. 백지원과 정연경(2006)은 현행의 용어관계가 가진 문제점을 파악하기 위하여 용어관계의 다양한 사례를 조사·분석하고, 이를 바탕으로 용어관계를 체계화하고자 하였다. 그 결과, 실재하는 용어관계의 유형의 수는 매우 많고 그 유형이 다양하며, 전형적인 구분인 동등관계, 계층관계, 연관관계 간의 구분조차도 명확하지 않을 뿐만 아니라, 각 용어관계 내에서의 하위 구분도 제각기 다르며 불분명한 경우도 많다는 사실을 밝혀냈다. 용어간의 관계표현은 시소러스의 품질을 높이는 것과 상관관계가 있다. 각 언어의 개념들을 충분히 설명할 수 있는 어휘들을 가지고 있다고 해도 모든 이용자가 동의할 수 있는 용어간의 관계를 표현하는 것이 시소러스의 구축에 있어 가장 중요하게 해결되어야 할 문제이다.

2.3 정보전문가 의견

현재의 다국어 시소러스 구축과 관련된 문제점을 파악하기 위해 한·일 양국의 다국어 시소러스 구축 경험이 있는 11인의 관련 전문가(한국인 6명, 일본인 5명)에게 간단한 인터뷰 형식으로 다국어 시소러스의 구축과 관련된 질문을 하였고, 그 응답 내용을 정리해보면 다음과 같다.

(표 1) 한국과 일본의 다국어 시소러스 구축 경험자 인터뷰 내용

질문	응답내용
다국어 시소러스 구축 시 주로 참고하는 것	<ul style="list-style-type: none"> - ISO 표준 - 기구축 시소러스 - 관련텍스트(기존연구포함)
참고한 정보원의 문제점	<ul style="list-style-type: none"> - 기구축 시소러스 <ul style="list-style-type: none"> • 대부분의 기구축 시소러스가 영·미·유럽어권 위주이다. • 시소러스마다 주제특정적이어서 구축기준이 범용적이지 못하다. - ISO 표준 <ul style="list-style-type: none"> • 오래되어서 현재의 정보이용환경을 반영하기 어렵다. • 영·미·유럽어권 위주의 설명들로 구성되어 있다. - 관련 텍스트 <ul style="list-style-type: none"> • 새로운 아이디어 확보에 유용하나 현실적으로 적용하기 어렵다. • ISO 표준과 함께 설계 지침 확보에 가장 유용하다
다국어 시소러스 구축 시 어려운 점	<ul style="list-style-type: none"> - 현실에 적용할 수 있는 지침이 없다. - 어휘 수집범위를 결정하는 것이 어렵다. - 등가어 처리가 까다롭다. - 어휘간의 관계 설정이 어렵다.
기구축된 다국어 시소러스의 문제점	<ul style="list-style-type: none"> - 어휘의 통제에만 집중되어 있다. - 실제 정보환경에서의 활용도가 낮다. - 일반 이용자가 사용하는 어휘를 포함하고 있지 않다. - 시소러스의 유지·갱신이 거의 이루어지지 않는다.
다국어 시소러스 구축을 위한 제안	<ul style="list-style-type: none"> - 일반 이용자가 쉽게 접근할 수 있어야 한다. - 실제 정보검색환경에서 응용될 수 있어야 한다. - 기존의 등가어 처리방식이나 어휘간의 관계 설정방식에서 벗어날 필요가 있다.

우선, 시소러스 구축 경험자들은 다국어 시소러스 구축에 있어 ISO 표준을 가장 많이 참조하고, 그 외에 기구축된 시소러스나 관련 문헌을 참조하는 것으로 나타났다. 그러나 참고 자료들은 오래 되었거나, 언어권이 다르거나, 현실적으로 동떨어지는 방법론을 적용하였다는 이유로 참고의 대상이 될 뿐, 실제적으로 경험적 판단에 근거해서 다국어 시소러스를 구축해온 것으로 조사결과 나타났다. 이것은 기존의 구축 지침들이 현실에 적용하기 어려운 측면들을 많이 내포하고 있다는 것으로 해석될 수 있다. 이러한 문제들은 다른 질문에 대한 응답에서도 찾아볼 수 있는데, 다국어 시소러스 구축 시 어려운 점을 묻는 질문에서도 역시 참

고할만한 지침을 찾는 것이 쉽지 않다는 응답이 있어, 다국어 시소러스를 구축하는 데 있어 현실적인 구축방법론에 대한 요구가 있다는 것을 확인할 수 있었다.

기구축된 다국어 시소러스의 문제점을 묻는 질문에서 가장 많은 전문가들이 실제 정보환경에서 다국어 시소러스의 활용도가 낮다는 것을 지적하였다. 오랜 시간 작업 끝에 다국어 시소러스를 완성해도 시소러스의 유지 및 관리가 부족한 현실이기 때문에 현실적으로 정보검색 환경에서 다국어 시소러스를 활용하는 것이 쉽지 않다는 지적이 있었다. 또한, 앞의 선행연구 분석에서도 언급하였듯이, 다국어 시소러스 구축 시 가장 중요하게 고려되는 두 가지 문제,

즉 등가어의 처리와 어휘간의 관계 설정이 어렵다는 문제점을 언급하였고, 이외에도 다국어 시소러스를 구축하고자 할 때 어휘를 수집하는데 있어 전통적으로 사전에 의지해왔던 방식 외에도 새로운 어휘수집방법을 적용할 필요가 있다는 의견이 다수 있었다.

이러한 문제점에 대한 언급은 ANSI/NISO에서 Z39.19 표준을 개정하기 전에 실시한 설문조사의 결과와도 어느 정도 일치하고 있다. ANSI/NISO에서는 Z39.19 표준을 개정하기 이전에 문헌정보학 분야의 온라인 커뮤니티의 이용자들을 대상으로 Z39.19의 개정과 관련한 설문조사를 실시하여 80개의 응답을 얻었다. 이 중 Z39.19 표준의 개정에 추가하고 싶은 내용에 대한 질문에 대해 응답자들은 온라인/웹 환경에서 이용이 가능한 시소러스를 설계할 수 있는 지침을 마련해줄 것과 동시에 온라인 탐색환경을 개선할 수 있도록 표준의 내용을 개정해줄 것을 요구하였다(ANSI/NISO 2003). 이러한 응답의 내용은 앞에서 언급한 정보전문가들의 관점과 유사한 것으로, 기존의 시소러스 구축지침이 현재의 정보환경에 응용될 수 있는 시소러스를 구축할 수 있는 지침으로서의 기능을 충분히 수행하지 못하고 있다는 것을 증명하고 있다.

2.4 다국어 시소러스 구축의 문제점

앞에서 언급된 내용을 중심으로 이용자 친화적 다국어 시소러스를 구축하는 데 있어 현재의 다국어 시소러스 구축방법을 적용하는 경우 제기될 수 있는 문제점을 분석하면 다음과 같다.

첫째, 참조할 수 있는 다국어 시소러스 지침에

문제가 있다. 앞에서 언급하였듯이 다국어 시소러스 구축 표준으로 전세계에서 이용되고 있는 ISO 5964는 이미 발표된지 20년이 지났다. 20여 년 전과 현재의 정보환경을 비교해보면 인터넷의 보급과 통신기술의 발달로 인한 도서관 환경의 변화 등 많은 변화가 있었음을 알 수 있다. ANSI/NISO의 Z39.19 표준 개정의 이유에서도 밝혔듯이, 과거의 시소러스 관련 표준들은 그 성격이 도서관 업무 중심적(library-centric)이고, 텍스트 문서 중심적(document-centric)이며, 인쇄자료 중심적(print-oriented)이다. 즉, 이 두 표준은 현재 정보유통 환경의 변화가 제대로 반영이 되지 않고 있으며, 이러한 문제는 결국 이용자 비친화적인(user-unfriendly) 다국어 정보유통을 초래한다.

둘째, 이용자들은 기구축된 다국어 시소러스를 현재 정보환경에서의 정보검색 보조도구로 이용하지 않고 있다. 이용자들은 시소러스와 같은 검색보조도구의 필요성은 느끼고 있으나 실제로 이용자들이 이용하기에 시소러스는 현실과 동떨어진 경향이 있다. 다국어 시소러스가 이용자 친화적인 형태가 되려면 우선적으로 인터넷 정보유통 환경에서 각 언어 이용자의 다원적인 정보이용 패턴을 분석할 필요가 있다.

셋째, 시소러스 구축방법론의 문제로 인해 다국어 시소러스가 대상 언어의 문화적 특성을 반영할 수 없다. 일반적으로 다국어 시소러스를 구축하는 방법에는 다음과 같은 세 가지가 있다(Hudon 1997).

- 1) 기존의 단일어 시소러스를 다른 언어로 번역하는 방법
- 2) 하나의 원문언어를 중심으로 기존의 단일

어 시소러스를 통합·조정하는 방법

- 3) 각 언어별 시소러스를 구축한 후 이들을 통합하고자 할 때 각 언어가 원문언어와 목적언어 역할을 모두 수행할 수 있도록 단일어 시소러스를 통합·조정하는 방법

1)과 2)의 방법은 목적언어의 특성을 제대로 반영할 수 없기 때문에 문제가 있으며, 3)의 방법이 가장 완벽한 다국어 시소러스를 구축할 수 있는 방법이나 구축시간이 오래 걸린다는 문제점이 있어 일반적으로 잘 사용되지 않는다.

일반적으로 많이 사용하는 방법은 2)번인데, 하나의 원문언어를 두고 기존의 단일어 시소러스 어휘들의 등가어를 처리하면서 통합하는 방식을 취하고 있다. 이러한 방법을 통해 시소러스를 구축할 때 생길 수 있는 오류는 다음과 같다(Hudon 1997).

- ① 원문언어가 갖는 개념구조에 맞도록 목적언어를 무리하게 끼워 넣어 화자가 이해할 수 없는 정도로 목적언어의 어휘가 손상될 수 있다.
- ② 원문언어의 문화적 배경을 목적언어에 그대로 강요하여 목적언어의 문화적 특성을 살릴 수 없다.
- ③ 목적언어를 문자 그대로 원문언어로 옮김으로써 목적언어의 언어적 의미를 상실할 수 있다.

이러한 방식으로 구축하는 경우는 원문언어 중심으로 통합이 이루어지기 때문에 목적언어의 언어적 특성을 반영하는 것이 어려우며, 이것은 대상 언어에서 표현되는 문화적 차이를

시소러스에 반영하기 쉽지 않다는 것을 의미한다. 다국어 시소러스의 구축은 단순히 여러 언어의 어휘들을 통합, 조정하는 데 그치는 것이 아니다. 언어라는 것은 그 언어가 사용되는 문화를 대변한다. 그렇기 때문에 이용자에게 효과적으로 이용될 수 있는 시소러스라면 그 언어의 문화까지도 반영할 수 있는 구조로 구축되어야 한다.

넷째, 시소러스 구축을 위한 어휘수집 단계에서의 문제이다. 기구축된 시소러스를 살펴보면 대부분 특정 주제분야(영역)를 중심으로 관련된 어휘를 대상으로 시소러스를 구축하고 있는데, 이 때 어휘수집방법으로 관련분야의 용어사전, 텍스트 등이 일반적으로 사용된다. 그러나 사전과 텍스트 등에 출현한 어휘는 이미 통제어휘(controlled vocabulary)의 성격이 강하다. 시소러스가 어휘통제를 주목적으로 색인 작성자들을 중심으로 사용되던 때에는 통제어휘를 수집하여 시소러스를 구축하는 것이 적절한 방법이었다. 그러나 오늘날과 같이 최종이용자가 인터넷을 통해 직접 정보를 검색하고 이용하는 정보이용 환경에서는 시소러스가 가지고 있는 어휘의 통제기능보다는 정보검색의 보조수단으로 이용되는 것이 시소러스의 효용성을 높이는 데에 더욱 긍정적인 효과를 나타낼 수 있다. 시소러스가 일반 이용자들에게 정보검색의 보조수단으로 활용될 수 있으려면 시소러스 구축에 필요한 어휘수집 역시 일반인들이 이해할 수 있는 자연언어 형태의 어휘들을 포함해야 한다.

다섯째, 시소러스의 용어간 관계표현에 관한 문제이다. 시소러스에서 용어와 개념의 관계는 일대일관계가 이상적이지만, 개념은 자연

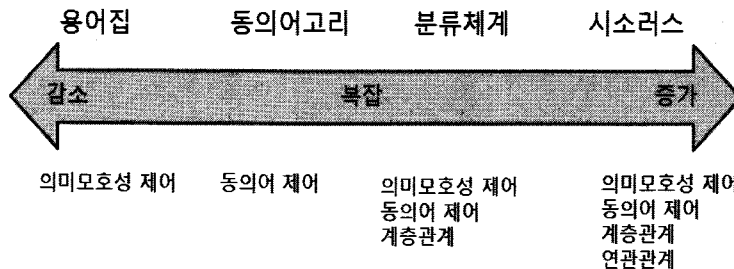
언어를 이용하여 나타내기 때문에 용어와 일대 일관계가 유지되지 않는 문제가 발생한다. 그럼에도 불구하고 지금까지 이것이 시소러스에서 그다지 문제시되지 않은 것은 시소러스에 의미를 다루려고 하는 의도가 없었기 때문이기 보다는, 시소러스의 용어가 다른 용어와의 관계 그 자체로 정의되거나, 시소러스가 전문용어를 다루고 있을 뿐 아니라 그 주제의 전문가들만 사용하기 때문에 의미를 동일하게 인식하고 있을 것이라 생각한 것으로 보인다(김지훈, 김태수 2006). 그림 3에서 보는 바와 같이, 통제어휘 중에서 시소러스가 갖는 구조적 복잡성은 의미모호성 제어, 동의어 제어, 계층관계, 연관관계 등으로 규정할 수 있지만, 이러한 구조적 복잡성을 가진 시소러스를 기존의 세 가지 관계(동등관계, 계층관계, 연관관계) 안에서 표현하는 것은 무리가 있다. 이러한 이유로 일부의 시소러스 연구에서는 용어간의 관계 설정은 연구의 범위에서 제외하고 단순히 언어간의 개념 일치화 과정에 중점을 두는 경향을 보이고 있다.

3. 인터넷 정보환경과 시소러스

3.1 이용자 태그와 시소러스

우리나라에서 기존의 다국어 시소러스 개발 양상을 살펴보면, 대부분 국내 관련분야 전문가의 의견 수렴만을 바탕으로 다국어 시소러스를 구축하였으며, 다국어 시소러스를 구성하는 각 언어별 시소러스의 실제 이용자 집단에 대한 정보이용 분석이 이루어지지 않고 있다. 그렇기 때문에 그 동안 개발된 다국어 시소러스는 국내의 소수의 이용자 집단만을 이용대상으로 규정해왔다. 결과적으로 언어별 쌍방의 정보유통을 가능하게 하는 시소러스가 아니라, 한국어를 원문언어로 두고 다른 외국어를 목적언어로 두는 일방적 정보유통을 위한 시소러스를 개발한 것이라 할 수 있다.

한편, 오늘날의 인터넷 정보환경은 웹 2.0 환경으로 정의할 수 있다. 이용자의 참여를 근간으로 하는 웹 2.0 환경은 다양한 방식으로 웹 이용자들이 정보의 생성과 공유에 직접 참여하도록 유도하고 있다. 이러한 웹 2.0 환경에서 시소러스의 구축과 관련하여 주목해야 할 내용이



(출처: ANSI/NISO Z39.19, p.17)

(그림 3) 통제어휘간의 구조적 복잡성의 상관관계

있는데, 그것은 바로 이용자가 정보에 대해 직접 그 내용을 표현하는 꼬리표(tag)를 달 수 있게 하는 '협력적 태깅(collaborative tagging)'이다. '이용자 태깅(user tagging)'이라고도 불리는 이 서비스는 이용자로 하여금 정보자원에 대해 요약할 수 있도록 일종의 색인어를 표시하게 함으로써 정보에 대한 내용을 쉽게 파악할 수 있도록 유도한다. 또한 특정 태그와 연관된 다른 태그들을 이용자간에 서로 공유할 수 있도록 하고 있어 이용자는 편리하게 정보자원을 관리할 수 있다. 특정 정보에 대해 이용자가 태그를 작성하는 행위는 전통적인 도서관 환경에서 마치 정보자원에 대해 시소러스 등과 같은 통제어휘집을 참조하여 색인어를 부여하는 행위와 유사하다.

태그작성에 사용되는 어휘들은 완전히 통제언어의 성격을 갖는다고 할 수도 없지만 그렇다고 완전히 자연언어의 성격을 갖는다고도 할 수 없다. 이용자가 평소에 사용하는 어휘로 정보의 내용을 표현한다는 관점에서는 자연언어의 성격을 갖는다고 할 수 있으나, 태그에 사용된 어휘들을 서로 공유하면서 이용자들은 다른 이용자들의 영향을 받아 알게 모르게 어휘의 통제작업을 거치게 된다. '어휘의 사회화'가 이루어지는 것이다. 그렇기 때문에 이용자 태그는 이용자가 직접 생산해낸 일종의 어휘집이라고 생각할 수 있다.

현재의 인터넷 정보환경에서 이용이 가능한 시소러스를 구축하기 위해서는 현재 인터넷에서 실제로 통용되고 있는 어휘들을 대상으로 해야 한다. 이러한 관점에서 이용자 태그를 시소러스 구축에 응용하는 방법을 생각해 볼 수 있다. 이와 관련해서 Spiteri(2007)는 태그가

통제어휘로서 사용될 가능성이 있는지에 대해 연구한 바 있다.

3.2 한국과 일본의 소셜 태그 분석

한국과 일본의 소셜 북마크 서비스인 'mar.gar.in(<http://mar.gar.in/>)'과 'hatena(<http://www.hatena.ne.jp/>)'를 대상으로 다섯 개의 태그(블로그, 웹2.0, 여행, 게임, 쇼핑)에 대한 연관태그를 분석하였다. 연관태그는 하나의 태그에 대해 많은 이용자가 동시에 태그로 작성한 것 중 동시출현빈도가 높은 태그들로 구성된다. mar.gar.in의 경우 특정 태그와 관련된 연관태그를 자동으로 제공해주고 있지만, hatena의 경우에는 연관태그를 자동으로 제공해주고 있지 않아, 분석대상 태그와 동시출현횟수가 5회 이상이 되는 태그를 연관태그로 추출하였다. 표 2에서는 mar.gar.in과 hatena 사이트에서 태그 '블로그'에 대해 제시된 연관태그의 예를 보여준다.

표 2에서 밑줄친 부분은 태그 '블로그'에 대해 양측 사이트에서 공통적으로 연관된 태그들을 의미한다. mar.gar.in에서 17개, hatena에서 13개의 연관태그를 얻었으나 이들 중에서 양측 모두 정확히 일치하는 태그의 수는 5개에 불과하다. 이러한 현상은 조사된 다른 태그들에서도 나타났는데(표 3), 심지어 '여행'과 '게임'에 대해서는 일치하는 연관태그 수가 0인 것으로 나타났다. 이것은 실제 정보환경에서 정보이용자들은 같은 개념에 대해 관심이 있다고 해도 이들의 인지구조 속에 포함되어 있는 그 개념과 관련된 어휘는 굉장히 다양하다는 것을 의미한다.

(표 2) hatena와 mar.gar.in 사이트에서 태그 '블로그'에 대한 연관태그

hatena	WEB, blog, WEB서비스(웹서비스), まとめ(정리해놓은 사이트), hatena, はてな(하테나), あとで読む(나중에읽을것), 読み物(읽을거리), communication, デザイン(디자인), webデザイン(웹디자인), WEB2.0, google
mar.gar.in	blog, IT, lang:ko, RSS, WEB, WEB2.0, 가져온즐거찾기, 검색, 구글, 네이버, 뉴스 디자인, 마가린, 메타블로그, 블로거, 사진, 웹2.0

(표 3) mar.gar.in과 hatena의 연관태그 수 비교

태그 \ 연관태그 수	mar.gar.in	hatena	공통
블로그	17	13	5
웹2.0	19	17	4
여행	2	5	0
게임	4	7	0
쇼핑	4	5	1

양측에서 공통된 태그를 찾아보기 어려운 이유는 이용자들의 지식구조 속에 다양한 어휘가 존재하기도 하지만, 언어가 문화적 차이를 반영하기 때문이라는 해석도 가능하다. 연관태그의 어휘들은 마치 단어의 연상작용에 의한 결과와 유사하다. 단어연상이란 어떤 단어가 주어졌을 때, 그것과 관련된 사항이 머리 속에 떠오르는 심리적 작용을 설명하는 용어로, 단어연상을 이용하면 특정 분야나 이용자 집단에서 나타나는 언어 및 어휘의 사용패턴을 확인할 수 있다(Nielsen 2001).

한승희(2006)는 단어연상검사법을 이용한 탐색 시소러스 구축과 관련된 연구에서 사람들의 인지구조의 차이에 따라 연상하는 어휘가 달라지기 때문에 이러한 어휘 중에서 자주 공통적으로 연상되는 어휘들을 모으면 시소러스 구축이 가능하다는 것을 실험을 통해 증명하였다. 그리고 기구축된 시소러스와 연상 시소러스를 대상으로 질의확장 실험을 통해 단어연상

을 통해 구축된 시소러스가 검색 성능에 있어 우위에 있다는 사실을 밝혀냈다. 이찬규(2002)와 Tsuji(2004)는 단어연상에는 문화기반마다 연상에 차이가 있다는 '문화적 연상(cultural association)'이라는 개념을 설명하였는데, 이러한 관점에서 표 3의 공통된 연관태그의 수가 적은 이유 역시 언어를 포함한 문화기반의 차이에 따라 특정 개념에 대해 연상할 수 있는 어휘가 달라질 수밖에 없는 것으로 해석할 수 있다.

4. 탐색 시소러스 개념을 이용한 이용자 친화적 한·일 시소러스 모형

다국어 시소러스가 현재 이용자의 정보환경에서 이용자 친화적인 형태로 구축되기 위해 갖추어야 할 조건은 다음과 같다.

첫째, 이용자가 정보이용 환경에서 실제로

사용하는 자연언어에 가까운 어휘로 시소러스를 구축해야 한다. 이러한 어휘로 시소러스를 구축하면 각 언어의 문화적 배경을 그대로 표현할 수 있어, 문화 게이트웨이(cultural gateway)로서의 시소러스의 기능을 수행할 수 있다.

둘째, 인터넷 정보검색 시의 검색보조도구로 이용되어야 한다.

이러한 두 가지 조건을 만족시킬 수 있는 다국어 시소러스의 모형으로 이 연구에서는 다국어를 지원하는 탐색 시소러스를 제안하고자 한다. 탐색 시소러스로 다국어 시소러스를 구축하는 이유는 이용자의 자연언어를 시소러스 구축에 반영할 수 있고 완성된 탐색 시소러스를 가지고 검색 성능을 향상시킬 수 있기 때문이다.

탐색 시소러스는 색인어의 제어를 목적으로 설계된 전통적인 시소러스와는 근본적으로 설계를 달리하고 있다. 즉, 미리 정해진 틀로 용어의 형식을 제어하는 데에는 관심을 두지 않고, 오히려 의미연상을 위해 동의어와 관련용어간의 연결을 중요시한다(김태수 2000).

탐색 시소러스의 개념은 온라인 정보검색 환경에서 최종이용자가 직접 원하는 정보를 검색하던 시기에 탐색자가 사용하는 어휘와 데이터베이스의 어휘를 연결하기 위한 목적으로 개발되었다(Bates 1986). 이용자가 지닌 고유한 배경과 경험으로 인해 하나의 개념을 동일한 용어로 표현하기 어렵고, 또 비록 같은 사람일지라도 시간이나 환경에 따라 동일 개념을 상이한 용어로 표현할 수 있다는 문제점을 해결하기 위한 방안으로 제시된 것이 바로 탐색 시소러스이다. 탐색 시소러스를 이용하면 탐색자는 연상작용에 따라 관심있는 주제와 관련된 주제를 탐색하고 관련 자료를 검색하기 위해 다수

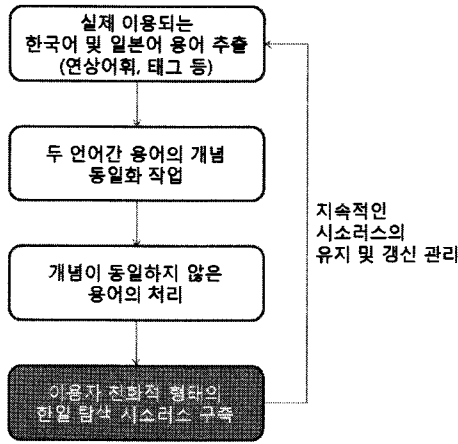
의 용어를 확인할 수 있다.

전통적인 시소러스와는 달리 이 탐색용 시소러스에서는 용어의 제어나 표준화를 시도하지 않으며 그 대신 탐색자가 생각하는 용어를 제시하고 관련된 개념을 연상할 수 있도록 다수의 연관어휘를 제시한다(김태수 2000). 이처럼 연상되는 용어간을 연결하거나 자극함으로써 탐색자의 지각과 인지범위를 확장할 수 있어 탐색전략을 아주 정교하게 구축할 수 있다. 이러한 탐색 시소러스는 인간의 인지적 특성에 의해 정보검색과정에서 생겨날 수 있는 불확실성(uncertainty), 다양성(variety), 복잡성(complexity)을 어느 정도 감소시킨다는 측면에서 정보검색 환경에 적절한 시소러스 모형으로 평가될 수 있다(Piternick 1984, Bates 1986).

그림 4에서는 탐색 시소러스의 개념을 적용하여 이용자 친화적인 한·일 시소러스 구축을 위한 모형을 제시하고 있다. 이 시소러스의 구축과정은 다음과 같이 설명될 수 있다.

첫째, 실제 이용되는 한국어 및 일본어 어휘를 추출한다. 이 때 자연언어에 가까운 어휘를 추출하기 위해 특정 개념에 대한 연상어휘 혹은 소셜 태그 등을 활용한다. 연상어휘는 검색 시에 이용자가 입력한 질의로그에서 획득할 수 있는데, 특정 개념과 함께 질의어로 입력하는 용어 중 일정 빈도 이상의 용어들을 시소러스 구축의 대상으로 포함시킨다.

둘째, 추출된 각 언어의 용어들을 대상으로 개념 매핑을 실시한다. 개념의 매핑은 완전일치를 우선으로 하는데, 이 때에는 『연세한국어사전』이나 『分類語彙表』와 같은 어휘집이나 일한대역사전을 활용한다.



(그림 4) 이용자 친화적 한·일 시소러스 구축을 위한 모형

셋째, 두번째 단계의 개념 매핑에서 제외된 불일치 용어를 처리한다. 다국어 시소러스 구축 과정 중 대부분의 개념 동일화 과정에서 야기된 문제점들은 문화적 배경의 차이 때문이다. 일반적으로 개념 동일화의 문제점은 문화적 차이, 언어적 차이, 관습적 차이에서 나타난다(남영준, 안동언 2000). 불일치 용어에 대한 매핑 작업은 양언어에 해박한 지식을 요하는 고도의 숙련작업이다.

불일치 용어를 처리하는 일반적인 방법은 다음과 같다. 우선, '수박 겉핥기'나 '빚질' 등과 같이 대응어가 없거나 정확한 개념과 매칭이 이루어지지 않는 경우에는 그 개념을 일역/한역하여 용어의 정의를 설명하는 것이 바람직하다. 또한 상대방 문화를 그대로 수용한다는 관점에서 번역을 하지 않고 그대로 사용하게 한 후 해당 용어가 보편화될 경우에 적당한 용어로 번역하도록 하는 방법도 가능하다. 언어적 차이로 완전히 일치하지 않는 경우에 있어서는 가장 보편적인 용어를 채택하여 의미를 부여하는 방법을 사용할 수 있다. 예를 들어 '다듬이질'은 한국

어에는 있지만 일본어에는 정확히 매치되는 용어가 없다. 이 때 가장 의미가 보편적인 'たたく(두들기기)'라는 용어로 번역해주는 것이 좋다. 또한 불일치 용어의 관리의 시소러스의 품질관리 측면에서 지속적으로 유지되어야 할 필요가 있다.

이러한 세 가지 단계를 거치고 나면 이용자 친화적 형태의 한·일 탐색 시소러스를 구축할 수 있다. 이 탐색 시소러스는 새로운 어휘를 쉽게 추가할 수 있기 때문에 시소러스의 지속적인 유지 및 갱신 관리가 가능하다.

5. 결론 및 제언

기존의 다국어 시소러스 구축 양상은 어휘의 '통제'를 중심으로 한 것이었으나, 정보환경의 변화에 따라 이러한 유형의 통제어휘집은 최종 이용자의 정보검색을 위한 보조도구로서의 기본 기능을 제대로 수행하지 못하고 있는 실정이다. 이 연구에서는 용어의 '통제' 뿐만 아니라

이용자의 '이용'까지 고려한 이용자 친화적인 다국어 시소러스를 구축하기 위한 모형을 제시하였다. 이를 위해 선행연구와 전문가와의 인터뷰를 통해 기존의 다국어 시소러스 구축법에서 제기될 수 있는 문제점을 분석하였으며, 현재의 정보환경에서 언어적 혹은 문화적 차이에 따른 이용자들의 어휘사용의 차이를 확인하기 위해 한국과 일본의 소셜 북마크 서비스를 대상으로 이용자 태그를 분석하였다. 본 연구에서 제안된 모형의 의의는 탐색 시소러스의 개념을 다국어 시소러스 구축에 적용하였다는 것

에 있다.

이 연구결과를 보다 일반화하기 위해 본 연구에서 제시된 모형을 이용하여 한·일 다국어 시소러스의 프로토타입을 구축하고, 구축된 다국어 탐색 시소러스가 갖는 탐색보조도구로서의 가능성을 확인하기 위해 질의확장실험을 수행할 필요가 있다. 뿐만 아니라, 한국어와 일본어에 대한 전반적인 언어적 이해를 토대로 한 개념 동일화 방법과 개념을 동일화 하지 못한 용어의 처리방법에 대한 연구가 필요하다.

참 고 문 헌

- 고은숙. 2007. 일본어와 한국어 형용사 대조 시소러스 연구. 『일본연구』, 31: 237-259.
- 國立國語研究所. 1996. 『分類語彙表』形式による語彙分類表. 増補版.
- 김지훈, 김태수. 2006. 용어정의와 관계추출을 통한 시소러스 확장에 관한 연구. 『한국문헌정보학회지』, 40(1): 293-314.
- 김태수. 2000. 『분류의 이해』. 서울: 문헌정보처리연구회.
- 남영준, 안동연. 2000. 多國語 시소러스의 概念 同一化 作業에 관한 實驗的 研究. 『지식처리연구』, 1(2): 57-78.
- 남영준. 2001. 지식관리시스템을 위한 의미형 한영 시소러스 구축에 관한 연구. 『한국도서관·정보학회지』, 32(4): 77-98.
- 『네이버 인조이재팬』.
<<http://enjoyjapan.naver.com/>>
- 백지원, 정연경. 2006. 용어관계의 분류 모형 개발에 관한 연구. 『정보관리학회지』, 23(1): 63-81.
- 백지원, 정연경. 2006. 지식조직체계의 용어관계 유형에 관한 연구. 『한국문헌정보학회지』, 39(4): 119-138.
- 연세대학교 언어정보개발연구원 편. 1998. 『연세한국어사전』. 서울: 두산동아.
- 이찬규. 2002. 단어연상에 관한 조사 연구(I). 『어문연구』, 30(2): 5-33.
- 이형대. 2001. 한국학 전문용어의 다국어 시소러스 구축 방법론. 『민족문화연구』, 35: 91-110.
- 日本電子化辭書研究所. 1995. 『EDR電子化辭書任様説明書』.
- 정영미, 김명옥, 이재운, 한승희, 유재복. 2002. 과학기술 분야 통합 개념체계의 구축 방안

- 연구. 『정보관리학회지』, 19(1): 135-161.
- 차재은. 2001. 어휘데이터베이스의 구축 및 활용 방안: 기존 어휘 자원의 활용과 다국어 DB로의 확장을 중심으로. 『민족문화연구』, 35: 73-90.
- 최석두, 조혜민. 다국어 시소러스의 설계. 『제8회 정보관리학회 학술대회 논문집』, 5-10.
- 한승희. 2004. 『클러스터링 기법을 이용한 지식구조의 자동 생성에 관한 연구』. 박사학위논문, 연세대학교 대학원 문헌정보학과.
- 한승희. 2006. 단어연상검사법을 이용한 탐색 시소러스 구축에 관한 실험적 연구. 『한국문헌정보학회지』, 40(3): 289-304.
- 한유석. 2000. 한국어 시소러스 작성의 방법과 문제점. 『한국일본어문학회 학술대회 발표논문집』, 17-21.
- 한유석. 2002. 다언어 시소러스 구축의 방법과 실제. 『日本語文學』, 12: 137-158.
- Aitichison, J., Gilchrist, A., & Bawden, D. 1997. *Thesaurus Construction and Use: A Practical Manual*. (3rd ed.) London: Aslib.
- ANSI/NISO Z39.19 *Summary of Survey Results*. [online]. [cited 2007. 10. 26]. <<http://www.niso.org/committees/T-RAG/Z39-19-Survey.pdf>>
- Bates, Marcia J. 1986. "Subject Access in Online Catalogs: A Design Model." *JASIS*, 37(6): 357-376.
- Chen, Jisong, Rowena Chau, and Chung-Hsing Yeh. 2003. "Discovering the Parallel Text from the World Wide Web." *The Australasian Workshop on Datamining and Web Intelligence*, 157-161.
- enjoykorea. <<http://bbs.enjoykorea.jp/>>
- hatena. <<http://www.hatena.ne.jp/>>
- Hudon, Michele. 1997. "Multilingual Thesaurus Construction: Integrating the Views of Different Cultures in One Gateway to Knowledge and Concepts." *Information Services & Use*, 17(2/3): 111-123.
- International Organization for Standardization. 1985. *Documentation - Guidelines for the Establishment and Development of Multilingual Thesauri (ISO 5964: 1985(E))*. Geneva: International Organization for Standardization.
- Maeda, Akira, Fatiha Sadat, Masatoshi Yoshikawa, and Shunsuke Uemura. 2000. "Query Term Disambiguation for Web Cross-Language Information Retrieval Using a Search Engine." *Proceedings of the 5th International Workshop Information Retrieval with Asian Languages*, 25-32.
- mar.gar.in. <<http://mar.gar.in/>>
- Nielsen, Marianne-Lykke. 2001. "A framework for work task based thesaurus design." *Journal of Documentation*, 57(6): 774-797.
- NISO. 2005. *ANSI/NISO Z39.19: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. Maryland: National In-

- formation Standards Organization.
- NISO. 2007. *Developing the Next Generation of Standards for Controlled Vocabularies and Thesauri*. [online]. [cited 2007. 10. 26].
<<http://www.niso.org/committees/MT-info.html>>
- Piternick, Anne B. 1984. "Searching Vocabularies: A Developing Category of Online Search Tools." *Online Review*, 8(5): 441-449.
- Spiteri, Louise F. 2007. "Structure and Form of Folksonomy Tags: The Road to the Public Library Catalogue: *Webology*, 4(2). [online]. [cited 2007. 10. 21].
<<http://www.webology.or/2007/v4n2/a41.html>>
- Tsujii, Y. 편. 임지룡, 요시모토 하지메, 이은미, 오카 도모유키 옮김. 2004. 『인지언어학 키워드 사전』. 서울: 한국문화사.
- Yang, Christopher C., and Kar Wing Li. 2003. "Automatic Construction of English/Chinese Parallel Corpora." *JASIST*, 54(8): 730-742.