다국어 시소러스의 설계*

Design of Multilingual Thesaurus

최석두, 이화여자대학교 문헌정보학과
sdchoi@mm.ewha.ac.kr
Suk-Doo Choi, Ewha Womans University

조혜민, 삼성의료원 의학정보센터
hmcho@smc.samsung.co.kr
Hye-Min Cho, Samsung Medical Center

둘 이상의 언어를 포함하고, 그 중 참조하는 언어를 중심으로 용어관계를 표시할 수 있는 다국어 시소러스의 개념구조에 대하여 논하였다. 아울러, 우리말을 기준어로 삼고, 기본요건, 구조화, 용어관계, 동형이의어, 표시방법, 배열 등에 대한 예시와 함께 다국어 시소러스의 설계방안을 제시하였다.

1 서론

대부분의 전문정보 색인 및 검색시스템에서는 정보를 담고 있는 매체와 기술되어 있는 언어를 색인항목으로 만들고 일괄해서 검색하거나 일부를 선택하여 검색할 수 있도록 하고 있다. 특히 다국어 정보검색에 대한 요구와 이를 충족시키려는 시스템들이 인터넷의 이용증가와 함께 더욱 증대하게 되었다. 그러나 다수의 언어를 동시에 검색하는 대부분의 시스템에서는 다음과 같은 문제점을 가지고 있다. 첫째, 동일한 의미를 갖는 검색어를 병렬로 나열해야 하는 번거로움이 있다. 둘째, 대응되는 외국어를 알아내야 하는 어려움이 있으며, 외국어의 종류가 매우 다양하다. 셋째, 중국어 및 일본어 문헌은 그 나라의 언어로 색인하고 검색하여야 하나 아직까지 그 처리가 원활하지 못하다. 넷째, 중국어 및 일본어를 대응되는 우리말 용어로 색인할 수 있는 어휘집이 없으며, 한자를 한글음으로 색인하게 되면 동형이의어가 많이 발생하게 된다.

이와 같은 문제점을 해결하는 방법은 다국어정보를 갖는 시소러스를 개발하는 것이다. 다국어 시소러스에 대한 연구 및 개발은 EC 등 유럽의 국제기관들을 중심으로 활발하게 진행되고 있으며 상용 소프트웨어도 등장하고 있다. 우리나라와 같이 선진국에 대한 학문의존도가 높은 경우 다국어 정보검색이 많이 요구되며 이에 따라 다국어 시소러스에 대한 연구 및 개발이 필수적이라 할 수 있을 것이다. 그러나 우리나라의 경우 문헌정보처리연구회에서 발행된 『시소러스 개발지침』에서 『*ISO 5964: 1985(E) Documentation - Guidelines for the Establishment and Development of Multilingual Thesauri*』가 번역 소개되었을 뿐 거의 연구되지 않는 실정이다. 따라서 본고에서는 다국어 시소러스 구축을 위한 일차적인 단계로 다국어 시소러스의 개념구조를 설계해 보고자 한다.

2 다국어 시소러스 개요

2.1 정의

다국어 정보검색 관련 논문들에서는 '다국어'라는 의미로 'multilingual', 'translingual', 'cross-lingual', 'cross-language' 등 뉘앙스가 다른 여러 종류의 용어

* 이 논문은 2001년도 두뇌한국21사업에 의하여 지원되었음.

를 혼용하여 왔다. 'Cross-Linguistic Information Retrieval' SIGIR-96 워크샵에서 참석자간에 'multilingual'이라는 넓은 의미보다는 보다 명확한 'cross-language'라는 용어로 통일하자는 의견이 있었지만, US DARPA(Defence Advanced Research Projects Agency)가 'translingual'을 사용함으로써 용어의 통일을 기하지 못하였다(Oard 1997).

마찬가지로 우리나라의 다국어 정보검색 관련 논문에서도 '교차언어', '다국어', '다언어'라는 여러 용어가 혼재하고 있다. 본고에서는 '다국어' 시소러스(multilingual thesaurus)로 사용하기로 한다.

ISO 5964는 "다국어 시소러스란 둘 이상의 언어로 구성된 시소러스로서 용어의 상호관계 및 각 언어의 등가어를 배열하고 있는 시소러스"라고 정의하고 있다. 이 정의에 따르면 다국어 시소러스는 두 가지 종류가 가능하게 된다.

하나는, 목적언어에 대하여 여러 언어로 대역어를 부기한 시소러스이다. 이 시소러스에서는 특정 용어를 참조하면, 참조용어의 용어관계를 목적언어로 보여주며, 각 용어에 대하여 대역어(대응어, 동의어)를 부기한다. 반대로 대역어를 참조하면, 그 대역어의 개념구조가 아니라 대역어의 대응어만을 보여준다. 또 하나는, 참조한 언어의 용어관계를 보여주며 각 용어에 대하여 대역어를 부기하는 것은 전자와 동일하지만, 대역어를 참조하면 참조한 대역어의 용어관계도 보여줄 수 있는 시소러스이다. 즉, 참조언어가 바뀔 수 있는 시소러스이다. 다만, 부기하는 언어를 지정하거나 디폴트 언어를 가질 수도 있다. 본고에서는 후자의 다국어 시소러스를 목표로 한다.

2.2 선행 연구

다국어 시소러스는 1960년대 정보검색에 대한 연구가 시작될 때부터 주목을 받아왔다. 최초의 다국어 시소러스로는 1964년 Pigur가 International Road Research Documentation System을 위해 개발한 영어, 불어, 독어 3개 국어 시소러스가 있다(Oard and Dorr 1996). 또한, Salton은 1970년 다국어 시소러스를 이용한 검색성능에 대해 보고하였다. Salton은 기존의 영어 개념리스트를 독일어로 번역하여 만든 다국어 개념리스트를 SMART 검색시스템에 적용하였다. 문헌정보학 관련 초록에 대해 48개의 영어질의를 수작업으로 독일어로 번역하고 실험하였다(Salton 1970).

OPAC에 관한 연구가 본격화된 1980년대 후반 Rolland-Thomas는 OPAC에서 다국어 전거화일의 필요성을 논하였다. 그는 이용자가 선택한 언어로 검색하였을 경우, 이용자의 언어에 매칭되는 자료뿐만 아니라 여러 언어로 된 전체 장서가 검색되어야 한다고 주장하였다(Rolland-Thomas 1989).

다국어 시소러스에 대한 본격적인 연구는 1990년대 중반 유럽지역을 중심으로 시작되었다. 대표적인 연구로는 1996년 시작된 EuroWordNet(EWN) 프로젝트를 들 수 있다. 이 프로젝트는 WordNet 1.5에 덴마크어, 독어, 이탈리아어, 스페인어, 불어, 체코어, 에스토니아어 등 유럽 7개 언어를 대상으로 단어간의 의미관계를 추가하여 대규모 다국어 시소러스를 구축하였다. EWN 데이터베이스로부터 4개 언어에 대한 대역어를 상호연결하는 중간언어 색인을 만들고 이를 이용하여 질의어와 문헌을 개념색인하여 개념기반의 문헌검색을 수행하고 있다(http://www.hum.uva.nl/~ewn/).

2.3 개발 사례

현재 다국어 시소러스가 구축되어 실제로 검색시스템에 이용되고 있는 대표적인 시소러스로는 농학분야의 AgroVoc, 환경분야의 GEMET, EC 활동과 관련된 분야를 다루는 EuroVoc 등을 들 수 있다.

AgroVoc은 1980년대 초 FAO와 EC 위원회에 의해 개발되었으며, 농학, 임학, 수산학, 음식, 환경 등의 관련분야에 대하여, 영어, 불어, 스페인어, 아라비아어 용어를 수록한 다국어 시소러스이다. 이 시소러스는 현재 AGRIS와 CARIS 정보시스템에서 사용되고 있다. AgroVoc의 1997년 3판에는 16,106개의 디스크립터, 9,450개의 영어 동의어, 8,693개의 불어 동의어, 12,086개의 스페인어 동의어 등, 도합 약 46,000 용어가 수록되어 있다. 이 시소러스는 인터넷을 통하여 이용할 수 있다.

European Environment Agent가 개발한 GEMET

(GEneral Multilingual Environmental Thesaurus)는 15개의 언어(영어, 네덜란드어, 핀란드어, 덴마크어, 독어, 노르웨이어, 스웨덴어, 프랑스어, 그리이스어, 이탈리아어, 포르투칼어, 스페인어, 헝가리어, 슬로바키아어, 미국 영어)를 지원하는 환경관련 다국어 시소러스이다. GEMET 2000년 판에는 5,298개의 디스크립터, 1,200개의 비디스크립터가 수록되었으며, 109개의 최상위용어, 1,264개의 영어 동의어가 수록되었다. GEMET는 THESmain으로 구축되고 관리되며 THES-show를 통해 일반인도 이용할 수 있다.

Eurovoc은 EC의 활동과 관련된 분야(정치, 국제관계, EC, 법률, 경제, 무역, 재정, 사회문제, 교육, 커뮤니케이션, 경영, 고용/노동문제, 수송, 환경, 농업/임업/수산업, 생산/기술/연구, 산업, 지리, 국제조직)를 다루는 다국어(독어, 영어, 네덜란드어, 스페인어, 핀란드어, 불어, 그리이스어, 이탈리아어, 포르투칼어, 스웨덴어 지원) 시소러스이다. 시소러스 구축 및 관리를 위해 TAT라는 시스템을 사용하고 있으며 인터넷을 통해 누구나 이용할 수 있다(http://europa.eu.int/celex/eurovoc).

한편 상용 시소러스 관리소프트웨어에서 다국어를 지원하는 경우가 있는데, 대표적인 소프트웨어로 THESmain과 MultiTes이 있다. THESmain은 ISO 2788과 ISO 5964의 기준에 따라 개발된 다국어 시소러스 관리시스템이며, Visual Basic 4.0과 MS Access를 이용하여 개발된 윈도우즈용 프로그램이다. 디스크립터, 비디스크립터, 관계 엔트리의 최대수는 저장공간만 있으면 제한이 없다. 시소러스에서 사용되는 언어는 30 종까지 가능하며 최대 100 종의 언어까지 확장이 가능하다. 외부 데이터베이스와 시소러스용어가 DDE-link로 연결가능하며 마이크로시소러스와의 합병도 용이하다.

MultiTes는 단독 PC나 LAN/WAN 환경에서 다양한 유형의 단어나 시소러스를 관리, 이용, 출력할 수 있는 다국어 시스러스 관리시스템이다. MultiTes는 terms, descriptors, top terms, subject headings, taxonomies, polyhierarchical thesauri, authority files, subject categories, multilingual thesauri를 생성할 수 있다. 시소러스 당 100만 용어를 구축할 수 있으며, 용어수, 계층의 깊이, 용어 당 관계의 수는 제한이 없다. 미국표준 ANSI/NISO Z39.19-199x *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*에 준거하고 있다.

3 다국어 시소러스 설계

3.1 기본 요건

다국어 시소러스의 기본기능은 단일어 시소러스와 동일해야 한다. 다만, 운용의 측면에서 고려해야 할 몇 가지 사항이 있다. 첫째, 수록할 언어에 해당하는 문자의 입력이 가능해야 한다. 특히 유럽특수문자, 일본식 약자 및 國字, 중국어의 簡體字 등의 표현과 입력이 편리해야 한다. 둘째, 용어의 다의성을 구분할 수 있어야 한다. 대상분야가 확대되면 2개 언어의 경우만 하더라도 다의성의 문제는 심각하며, 언어가 늘어남에 따라 그 수효와 관계가 방대하게 늘어나기 때문이다. 셋째, 기본적으로 우리말을 중심으로 하는 다국어 시소러스를 구축하여야 한다. 즉, 각 언어에 대한 우리말 대응어가 필수적으로 존재해야 한다. 구심점이 되는 우리말이 없다면 세상의 다국어 용어를 모아 놓은 시소러스가 되어 버린다. 다만, 시스템의 일시적인 표시기능으로 우리말 대신에 다른 언어를 표시할 수는 있다. 넷째, 용어의 분류가 필요하다. 단일어 시소러스에 비해 다국어 시소러스는 다의성이 높아진다. 이를 보완하는 방법 중의 하나가 용어분류이다. 여러 가지 분류방법이 사용되고 제안되고 있으나 대부분 매크로-다국어 시소러스 분류체계로는 부적합하다고 사료된다. 그중 가까운 분류체계로는 325,000여 어를 수록하고 분류하고 있는 Chapman (1992)의 *Roget's International Thesaurus* 분류체계(상위분류 15항목, 그 아래 하위분류 1073항목)를 들 수 있다.

3.2 언어간의 관계

용어간의 관계는 단일어 시소러스와 동일해야 한다. 다만, 용어관계 표현의 문제가 남아 있다. 원칙적으로 용어간의 관계는 모두가 같은 수준, 즉 등가어이다.

그러나 특정 언어를 기준어로 삼아 이를 중심으로 표현하는 것이 적용에 편리하다고 생각된다. 즉, 해당 언어 내에서는 그 언어로 정의된 용어관계만을 보여주되, 기준어를 항상 보여준다. 예를 들어, 우리말을 기준어로 지정하면, 우리말, 영어, 불어, 스페인어를 수록하고 있는 다국어 시소러스에서 우리말은 어느 언어를 선택하든 함께 보이게 된다. 각 용어에 대응되는 다른 나라 언어는 등가어로 처리한다. 후술하는 예에서는 우리말을 기준어로 삼지만, 특정 언어를 기준어로 지정할 수 있도록 시스템을 설계하면 필요에 따라 기준어를 바꿀 수 있을 것이다.

전형적인 예로, 다음 예1과 같은 우리말 시소러스에 대하여, 영어, 불어, 스페인어 용어를 갖는 다국어 시소러스를 만들고, 우리말을 기준어로 삼아 각 언어로 참조하여 보면 표 1과 같다.

예1) 열대 과일 [熱帶--]
BT 과일
NT 그레이프후르츠
망고
바나나
수박
오렌지
코코넛
파인애플

표 1은 우리말을 기준어로 삼고 영어를 대응어로 선택했을 때의 가장 단순한 관계표시이다. 따라서 대응어로 불어를 선택하게 되면 어느 언어를 참조하더라도 대응어는 언제나 불어가 되어야 한다.

참조용어로 지정하는 용어의 관련용어만을 표시할 수도 있지만, 이용자인터페이스의 설계에 따라서는 표 1의 내용 전체표시를 디폴트로 할 수도 있다. 예를 들면, '열대과일'을 참조하면, 이 용어에 대응되는 모든 용어, 즉 우리말 '열대과일', 영어 'tropical fruits', 스페인어 'frutas tropicales', 불어 'fruits tropicaux'에 대한 용어관계 모두를 함께 표시함으로써 타언어의 대응용어를 지정할 필요가 없게 된다. 다만, 관련용어나 언어의 수가 많아지게 되면 표시방법을 수정해야 할 필요가 있다.

표 1 선택언어별 용어관계 표시

한글 선택	열대 과일 [熱帶--] EN tropical fruits ES frutas tropicales FR fruits tropicaux BT 과일 / fruits NT 그레이프후르츠 / grapefruit 망고 / mango 바나나 / plantain 수박 / watermelon 오렌지 / orange 코코넛 / coconut 파인애플 / pineapple
영어 선택	tropical fruits / 열대과일 ES frutas tropicales FR fruits tropicaux BT fruits / 과일 NT coconut / 코코넛 grapefruit / 그레이프후르츠 mango / 망고 orange / 오렌지 pineapple / 파인애플 plantain / 바나나 watermelon / 수박
불어 선택	fruits toropicaux / 열대과일 EN tropical fruits ES frutas tropicales BT fruits / 과일 NT ananas / 파인애플 bananier / 바나나 mangue / 망고 noix de coco / 코코넛 orange / 오렌지 pamplemousse/ 그레이프후르츠 pasteque / 수박
스페인어 선택	frutas tropicales / 열대과일 EN tropical fruits FR fruits tropicaux BT frutas / 과일 NT coco / 코코넛 mango / 망고 naranja / 오렌지 pina / 파인애플 platano 바나나 sandia / 수박 toronja / 그레이프후르츠

3.3 다의성의 해소

표 1의 예와 같이 모든 용어가 정확하게 1대 1로 대응된다면 문제는 없다. 그러나 모든 언어는 원래의 의미이든 잘못 이해되든 다의성을 가지고 있다. 예를 들어, 우리말의 '기본권'에 대응되는 영어, 불어, 독어를 보면 다음과 같다[1].

기본권 [基本權]
ENG **civil rights**
fundamental human rights
fundamental rights
human rights
FRA droits de l'homme
GER fundamentale Menschenrechte
Grundrecht
Menschenrechte
Stammrecht

이 때 'civil rights'에 대응되는 우리말 '기본권' 이외에도 '공권(公權), 공민권(公民權), 민권(民權)' 등이 있다. 그러나 우리말에서의 세 가지 용어는 서로 다르다. 이들과 관련된 영어, 불어, 독어를 보면 다음과 같다.

공권 [公權]
ENG **civil rights**
public authority
GER burgerliche Ehrenrechte
offentliches Recht
공민권 [公民權]
ENG citizenship
civil rights
FRA droit civique
GER burgerliche Ehrenrechte
Burgerrecht
Staatsburgerrecht
민권 [民權]
ENG **civil rights**
people's rights
FRA droit du peuple
GER Rechte des Volkes

1) 필자는 매크로 한글시소러스를 구축하고 있으며, 현재의 수록용어는 약 220,000 어이다. 본 예는 이 시소러스에서 추출한 것이며, 대역어가 완전하다고 검증된 것은 아니다.

따라서 우리말을 참조했을 때는 문제가 없지만 'civil rights'를 참조했을 때, '공권, 공민권, 기본권, 민권' 중 어느 용어를 대응어로 삼을지 어렵게 된다. 이것은 각 언어가 갖는 공통적인 현상이다.

원칙적으로 이들을 구별할 필요가 있다. 구분방법은 한정어를 사용하는 방법이다. 한정어의 언어를 각각의 언어를 사용하는 방법과 우리말을 사용하는 방법이 있으나 우리말을 사용하는 것이 편리할 것이다. 그러나 한정어를 사용하더라도 '공권, 공민권, 기본권, 민권'과 같이 동일 意味族을 구분한다는 것은 매우 어렵다. 따라서 서로 완전히 상이한 의미를 갖는 경우에만 구별용 한정어를 사용하고, 동일 意味族인 경우에는 동일한 한정어를 부여하여 참조시 모두를 함께 나열하는 것이 바람직 할 것이다.

3.4 시소러스의 배열

ISO 2788에서는 일반 시소러스의 배열방법으로 용어 상호간의 관계에 대한 범위주기와 지시어를 가진 알파벳순배열, 알파벳순 색인에 의해 보완된 체계배열, 알파벳순 색인을 포함하는 도식배열의 세 종류를 제시하고 있다. 다국어 시소러스에서도 기본적으로 이 배열 체계를 따른다. 또한, ISO 5964에서는 우선어이든 비우선어이든 각 언어별 알파벳순 배열을 제시하고 있다. 즉, 모든 언어의 통합배열에는 의미를 두지 않고 있다.

책자형인 경우에는 대응어를 부기하는 언어별 알파벳순 배열(즉, 2개 국어 시소러스)이 많았으며, 경우에 따라 컬럼별로 다른 언어를 배열하는 경우도 있었다. 이 방법은 이용자가 여러 언어의 내용을 동시에 살펴볼 수 있다는 장점은 있으나 각각의 기준언어별로 중복 수록되므로 많은 공간이 소요되게 된다. 컴퓨터로 처리하여 모니터 상에 표시하는 경우에는 복수의 컬럼이나 복수의 窓을 만들어 여러 언어를 동시에 표시할 수 있다.

기본적인 배열의 사례로 표 1의 내용을 불어를 중심으로 배열하면 표 2와 같다.

4 결 론

다국어 시소러스를 설계하고자 할 때 기본적으로

고려해야할 구조에 대하여 논하였다. 아울러, 다국어 시소러스의 기본요건, 언어간의 관계, 다의성의 문제와 해결방안, 배열문제 등에 대하여 예시와 함께 기본 방안을 제시하였다.

표 2 용어의 자모순배열(불어를 기준한 경우)

ananas / 파인애플
EN pineapple
ES pina
BT fruits tropicaux / 열대과일
bananier / 바나나
EN plantain
ES platano
BT fruits tropicaux / 열대과일
fruits toropicaux / 열대과일
EN tropical fruits
ES frutas tropicales
NT ananas / 파인애플
bananier / 바나나
mangue / 망고
noix de coco / 코코넛
orange / 오랜지
pamplemousse / 그래이프후르츠
pasteque / 수박
mangue / 망고
EN mango
ES mango
BT fruits tropicaux / 열대과일
noix de coco / 코코넛
EN coconut
ES coco
BT fruits tropicaux / 열대과일
orange / 오랜지
EN orange
ES naranja
BT fruits tropicaux / 열대과일
pamplemousse / 그래이프후르츠
EN grapefruit
ES toronja
BT fruits tropicaux / 열대과일
pasteque / 수박
EN watermelon
ES sandia
BT fruits tropicaux / 열대과일

그러나 전술한 문제들도 모두 해결된 것은 아니며, 우리말을 중심으로 하는 다국어 시소러스의 실제 개발단계에서는 보다 다양한 문제들이 나타날 것이다. 시행착오를 거치지 않기 위해서는 보다 실제적이고 심층적인 연구가 필요하다고 사료된다.

참고문헌

문헌정보처리연구회. 1994. 『시소러스 개발 지침』. 서울: 동연구회.

ANSI/NISO Z39.19-199x *Guidelines for the Construction, Format, and Management of Monolingual Thesauri.*

Chapman, Robert L. ed. 1992. *Roget's International Thesaurus.* 5th ed. New York: Harper Collins.

Hudon, Michele. 1997. "Multilingual thesaurus construction-integrating the views of different cultures in one gateway to knowledge and concepts." *Information Services & Use,* 17: 111-123.

ISO 2788:1986(E). *Documentation - Guidelines for the Establishment and Development of Monolingual Thesauri.* Geneva: ISO.

ISO 5964:1985(E). *Documentation - Guidelines for the Establishment and Development of Multilingual Thesauri.* Geneva: ISO.

Oard, Douglas W. 1997. "Alternative approaches for cross-language text retrieval." In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence. <http://www.clis.umd.edu/dlrg/filter/sss/papers/oard.ps>

Oard, Douglas W. and Bonnie J. Dorr. 1996. *A Survey of Multilingual Text Retrieval.* <http://www.ee.umd.edu/medlab/filter/papers/mlir.ps>

Salton, Gerard. 1970. "Automatic processing of foreign language document." *Journal of the American Society for Information Science,* 21:187-194.

Soergel, Dagobert. 1997. "Multilingual thesauri in cross-language text and speech retrieval." In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence. <http://www.ee.umd.edu/medlab/filter/sss/papers/soergel.ps>.