우리말 시소러스 作成에 관한 硏究

A Study on Constructing Korean Language Thesaurus

金　泰　中*
(Kim, Tae-Jung)

抄　　錄

情報檢索시스템에서 統制語彙는 재현율을 높이고 索引者 또는 利用者가 적합한 用語를 선정하는데 도움을 준다. 시소러스는 統制語彙集의 한 형태로 대부분의 데이터베이스 製作者들이 사용하고 있다. 이 硏究의 目的은 우리말 시소러스의 作成方法을 開發하는 것이며 다음과 같은 內容을 다루었다. 1) 시소러스의 定義, 2) 시소러스 作成理論에 관한 文獻調査와 檢討, 3) 실제적인 시소러스 作成方法 提示, 4) 시소러스의 出力形態, 5) 實驗 및 實驗結果

ABSTRACT

In information storage and retrieval system, controlled vocabularies are generally used to improve the recall ratio and to guide for indexers/users to select correct indexing/searching terms by regulating their forms as well as meanings. Thesaurs, a type of controlled vocabulary, nowdays accepted by most of database producers.

The objective of this study is to develope a method of Korean Language Thesaurus construction. This study covers 1) the definition of thesaurus, 2) a literatural survey on term relations and thesaurus construction method, 3) a suggestion for a practical construction method, 4) the display format of thesuari, and 5) tests and the results.

* 産業硏究院 附設 産業技術情報센터 情報處理室.

Ⅰ. 序　　論

科學技術 특히 電子, 컴퓨터, 通信技術 그리고 交通의 발달로 情報化社會가 촉진되어 情報의 發生量은 폭발적이라고 표현할 만큼 급격한 增加趨勢를 보이고 있다. 예를 들어 1985 년 한해 동안 全世界의 科學技術分野에서 발표된 情報의 양은 3,696,000 件 [1] 에 이르고 있으며, 美國化學會의 CAS(Chemical Abstracts Service)가 발행하는 化學分野의 抄錄誌 *Chemical Abstracts* 誌에 수록된 抄錄數의 동향을 보면, 1959 년에 125,000 件, 1969 년에 252,320 件 [2] 이었으나 1988 년에는 474,545 件 [3] 이 수록되어 있다. 이와 같이 많은 情報 가운데 必要情報를 찾기란 쉬운 일이 아니기 때문에 실제에서는 情報의 內容을 분석하여 선택적으로 입수하게 되며, 입수된 情報는 抄錄, 分類, 索引 등의 과정을 거쳐 가공하게 된다. [4]

情報의 利用者는 목적에 따라 다양하고 특정적인 情報를 필요로 하므로 여러 관점에서 정보원에 접근할 수 있도록 되어야 하며, 이는 索引을 통해서 가능하다. 索引에서 가장 문제가 되는 부분은 索引言語이며, 索引言語는 著者의 文獻에 있는 主要 用語를 索引語로 사용하는 自然語(natural indexing language)와 索引에 사용되는 用語를 미리 정하여 사용하는 統制語 (controlled indexing language),그리고 索引作成者가 임의로 주제에 적합하다고 생각하는 索引語를 부여하는 自由 索引語(free indexing language)로 나누어지며, 自由索引語와 自然語는 索引作成時에 부여되는 用語의 다양성과 임의성으로 索引 본래의 목적을 충분히 이루기 위해 統制語에 보다 많은 관심이 집중되고 있다. [5]

文獻情報를 우리말로 처리함에 있어서 歐美의 경우보다 사용되는 用語의 統制가 더욱 절실하며 歐美의 경우는 대체로 복수와 단수 그리고 複合語의 처리에 관한 문제가 주로 논의의 대상이 되고 있으나, 우리말의 경우 단수와 복수는 문제가 되지 않지만 單語의 띄어 쓰기와 外來語의 表記 등이 보다 큰 문제이다. 또한 이러한 외형적인 통일 이외에도 사용되는 用語의 의미에 관한 定義 등에서도 統制가 필요하다.

이러한 의미에서 用語의 統制 및 標準化와 더불어 효율적인 情報處理 및 檢索을 위한 수단으로 시소러스가 가장 편리한 것으로 평가되고 [6] 있으며, 情報를 능률적으로 처리하고 檢索하는데 필요한 수단임에도 불구하고 적합한 우리말 시소러스가 없는 상태이다. 시소러스의 중요성을 알고 있으나 作成에 많은 費用과 時間이

소요되어 이 硏究를 통해 情報가 축적 안된 상태에서 컴퓨터를 이용하여 효율적으로 시소러스를 作成하는 方法을 강구하고자 한다.

Ⅱ. 理論的 背景

1. 시소러스의 定義

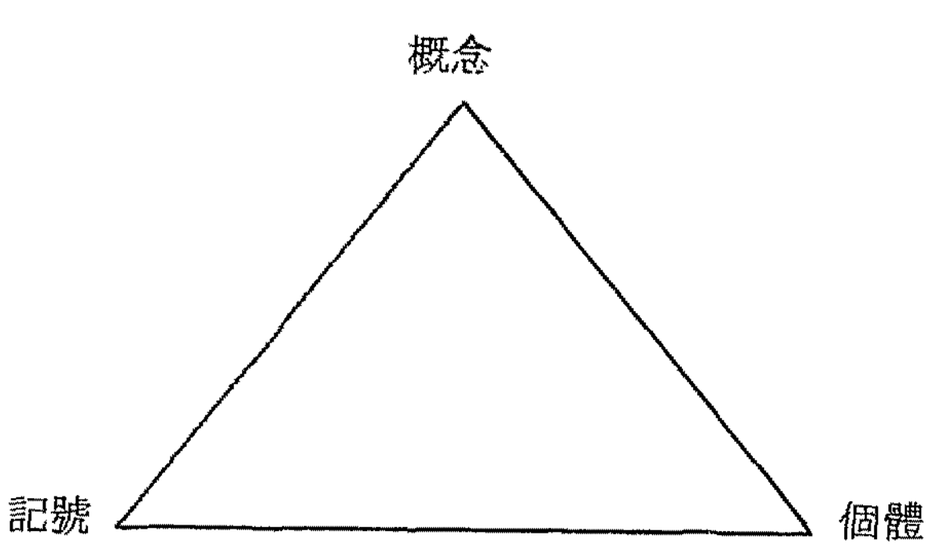
시소러스에 관한 定義는 다양하다. 그러나 ISO의 定義가 현대 시소러스의 機能과 構造를 가장 적절하게 표현하고 있다.[7] ISO는 시소러스에 관하여 "구조적인 면에서는 상위 개념의 用語와 하위 개념의 用語를 의미론적으로 밝힌 語彙集이며, 機能的인 면에서는 索引者 또는 情報利用者가 自然語를 統制語로 변환하는데 사용되는 語彙集이다"라고 정의하고 있다. 좀 더 구체적으로 풀어 본다면, 索引 및 檢索에 쓰이는 語彙를 표준화하고 情報의 축적 및 檢索시스템에서 同義語,上位槪念語, 下位槪念語, 關聯語 등의 선정이 용이하도록 각 用語의 개념적 관계를 밝힌 統制語彙集이다.

2. 槪 念 關 係

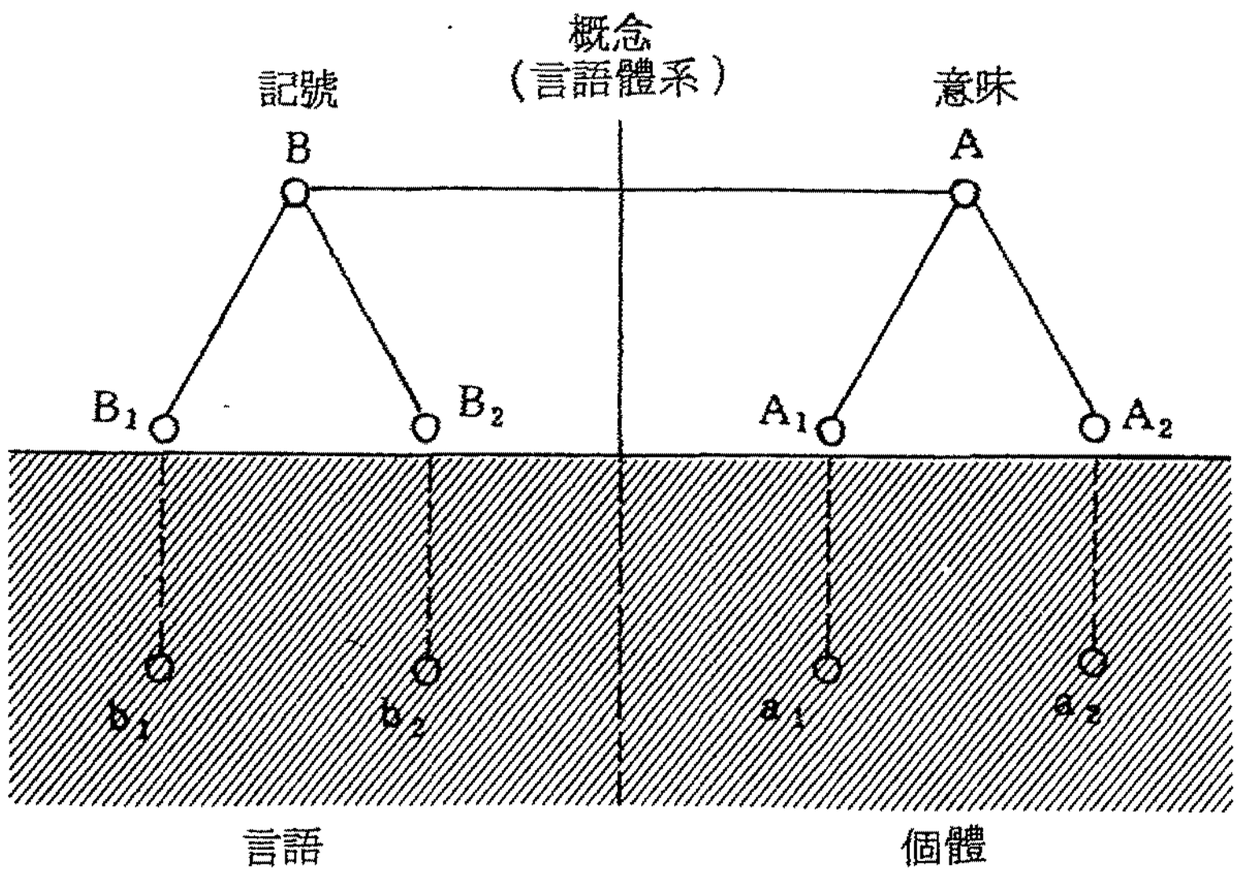
(1) 用 語 모 델

用語는 하나의 주제분야에서 定義된 槪念에 대한 관습적인 표상이며, 1908 년에 Gomperz 가 제안한 삼각형 모델 (〈圖 1〉)을 바탕으로 〈圖 2〉와 같은 모델을 1959 ~ 1960 년에 Wüster 가 제안하였다.[8) 9)]

〈圖 1〉 Gomperz 모델



〈 圖 2 〉 Wüster 모델



註 : B ····· 該當記號의 概念
B_1B_2 ····· 音聲 또는 記錄形態의 各概念
b_1, b_2 ····· 個體의 音聲 및 記錄形態
a_1, a_2 ····· 같은 種類의 個體
A ····· 概念 A_1과 A_2의 一般的 概念

(2) 概念間의 關係

시소러스에서는 同義關係, 階層關係, 關聯關係 등의 3가지 기본적인 관계가 用語間의 關係를 나타내는데 쓰이고 있다.

Wüster가 1971년에 발표한 概念間의 關係에는 2가지의 基本形態, 즉 概念分類를 위한 概念關係 (concept relation)와 시소러스와 같은 主題分類를 위한 主題關係 (subject relation)가 있으며, 概念關係는 〈表 1〉로 표현된다. Wüster는 主題關係는 概念關係에서 논리적 관계와 存在論的 關係의 구별이 필요없고, 階層關係만 존재하며 原料-製品關係, 時間關係, 因果關係, 作用關係, 系統關係 등에서는 關聯關係가 있다고 하였다.[10]

(3) 關 係 記 號

시소러스에 따라 關係 表示記號를 〈表 2〉와 같이 달리 사용하고 있으며[11], 보편적으로 USE, UF, BT, NT, RT, SEE, ALSO 등의 6가지 記號를 주로 사용하고 있다.

① USE, UF -同義 또는 類似關係에 사용한다.

USE는 비디스크립터를 디스크립터로 안내하며, UF는 디스크립터와 同義 또는 類似關係인 비디스크립터를 알려 준다. UF는 "Use for"의 줄인 말이다.

〈表 1〉 概念間의 關係

關係의 方向 / 關係形態	垂直 上位	垂直 下位	重複 重複	重複 同位	對角 對角關係
1 論理(抽象)關係	直接的 抽象關係 從屬關係		間接的 抽象關係 從屬의 殘餘關係		
	類(GENUS) 예) 과일	種(SPECIES) 사과	抽象的 重複	抽象的 同位	抽象的 對角關係
2 存在論的關係	直接的 存在論的 關係 部分關係		間接的 存在論的 關係 部分의 殘餘關係		
	部分的 上位 예) 전체	部分的 下位 부분 / 안구	部分的 重複	部分的 同位	部分的 對角關係
3 原料－製品關係	原料 製品 關係 예) 강 →	강관			
4 時間關係	繼承 예) 先行者 →	繼承者		同時性 共生	
5 5.1 因果	原因 →	效果			
5.2 作用	工具 →	工作			
5.3 디센트 (DESCENT)	5.3.1 系統的 예) 아버지 →	아들		형제 자매	삼촌조카
	5.3.2 發生學的 예) 계란 →	병아리			
	5.3.3 物質 變化 段階 (1) 原 油 → 揮發油 (2) 우라늄Ⅰ → 우라늄Ⅱ 우라늄Ⅱ → 라 듐 라 듐 → 라 돈				

〈表 2〉 시소러스에서 使用되는 記號

關係 \ 시소러스	TEST[12]	NASA THESAURUS[13]	科學技術用語 시스템[14]	ROOT[15]
同義	USE	USF	USE	→
同義	UF(Used For)	UF	UF	=
上位	BT(Broader Term)	GS (GENERIC STRUCTURE)	BT	〈
下位	NT(Narrower Term)		NT	〉
關聯	RT(Related Term)	RT	RT	－

② BT, NT－디스크립터간의 階層關係 즉, 上位 또는 下位槪念의 관계를 표시하며, BT는 "Broader Terminology"를 NT는 "Narrower Terminology"를 줄인 말이다.

③ RT－"Related Terminology"의 줄인 말이며, Wüster의 槪念關係에서 原料－製品關係, 時間關係, 因果關係, 作用關係, 系統關係 등을 표시하는데 사용된다.

④ SEE ALSO－디스크립터로 채택된 동의 또는 類似關係의 用語를 알려준다.

3. 作成方法論

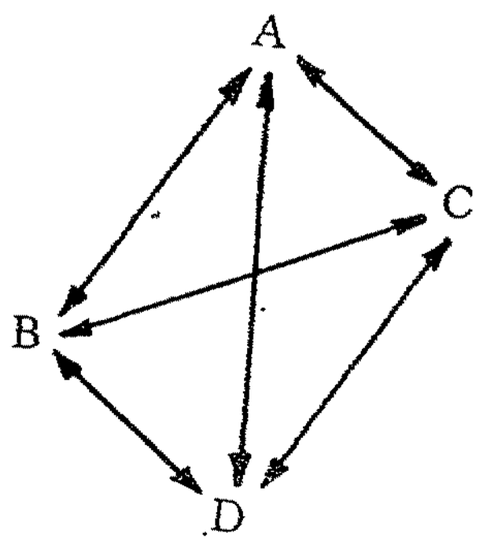
시소러스를 작성하는 데는 用語를 수집한 후에 槪念을 정립하는 方法과 槪念을 정립한 후에 用語 수집하는 방법으로 나누어진다.[16] 전자는 실제 쓰이고 있는 單語를 수집한 후에 수집된 용어 또는 用語間의 槪念 및 關係를 정의하는 귀납적 作成方法이며, 후자는 학문분야별 專門家들로 委員會를 구성하고, 委員會에서 分野別 槪念과 채택하고자 하는 用語를 정해진 槪念의 틀에 맞추는 연역적 作成方法과 같다.[17]

한편, 귀납적 방법은 현실성은 있으나, 주제의 포괄성이 결여될 우려가 있고, 연역적 방법은 主題의 포괄성에는 문제가 없으나 준비된 틀에 맞추기 위해 실제 흔히 사용되지 않는 用語가 수록될 여지가 많은 등의 문제점이 있어 ISO, BS 등의 規格에서는 두가지 方法을 혼합한 作成方法을 권장하고 있다.[18]

4. 用語 클러스터링 (Term Clustering)

관련이 있는 用語群은 文獻內에서 함께 쓰는 일이 많다는 가정하에 하나의 특정 주제를 다룬 文獻內에서 함께 자주 사용되는 用語들은 서로 관계가 있으며,[19] Salton은 다음과 같은 4가지 形態의 用語 클러스터 模型을 제안했다.[20]

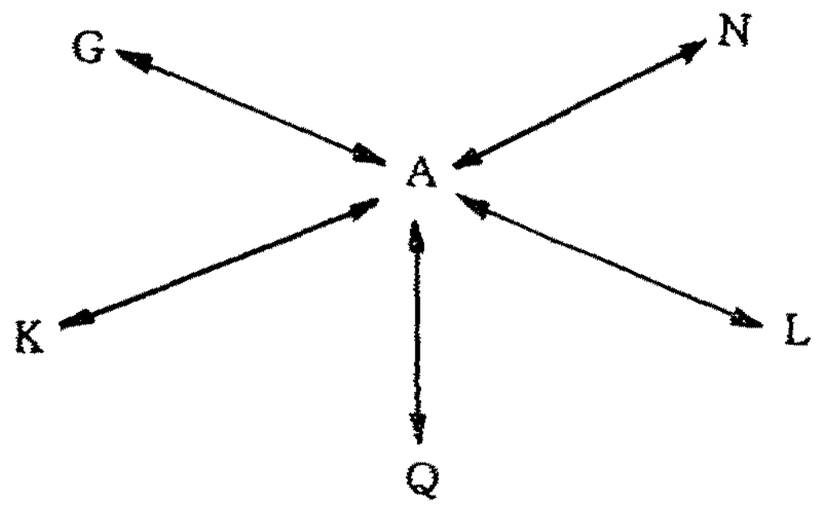
(1) 클리크 (Clique) : 항상 함께 쓰이는 用語群



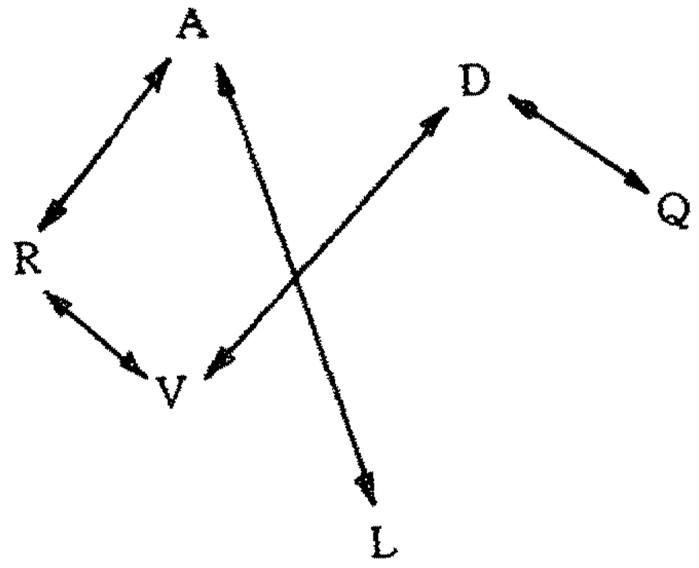
(2) 스트링 (String) : 用語들이 한쌍씩 쓰여서 용어 체인을 형성

A→D→G→L→P→Y

(3) 스타 (Star) : 여러 用語가 하나의 用語와 함께 쓰이는 용어군



(4) 클럼프 (Clump) : 특정한 성질을 보이지 않으나, 서로 이어지는 모양을 보이는 用語群

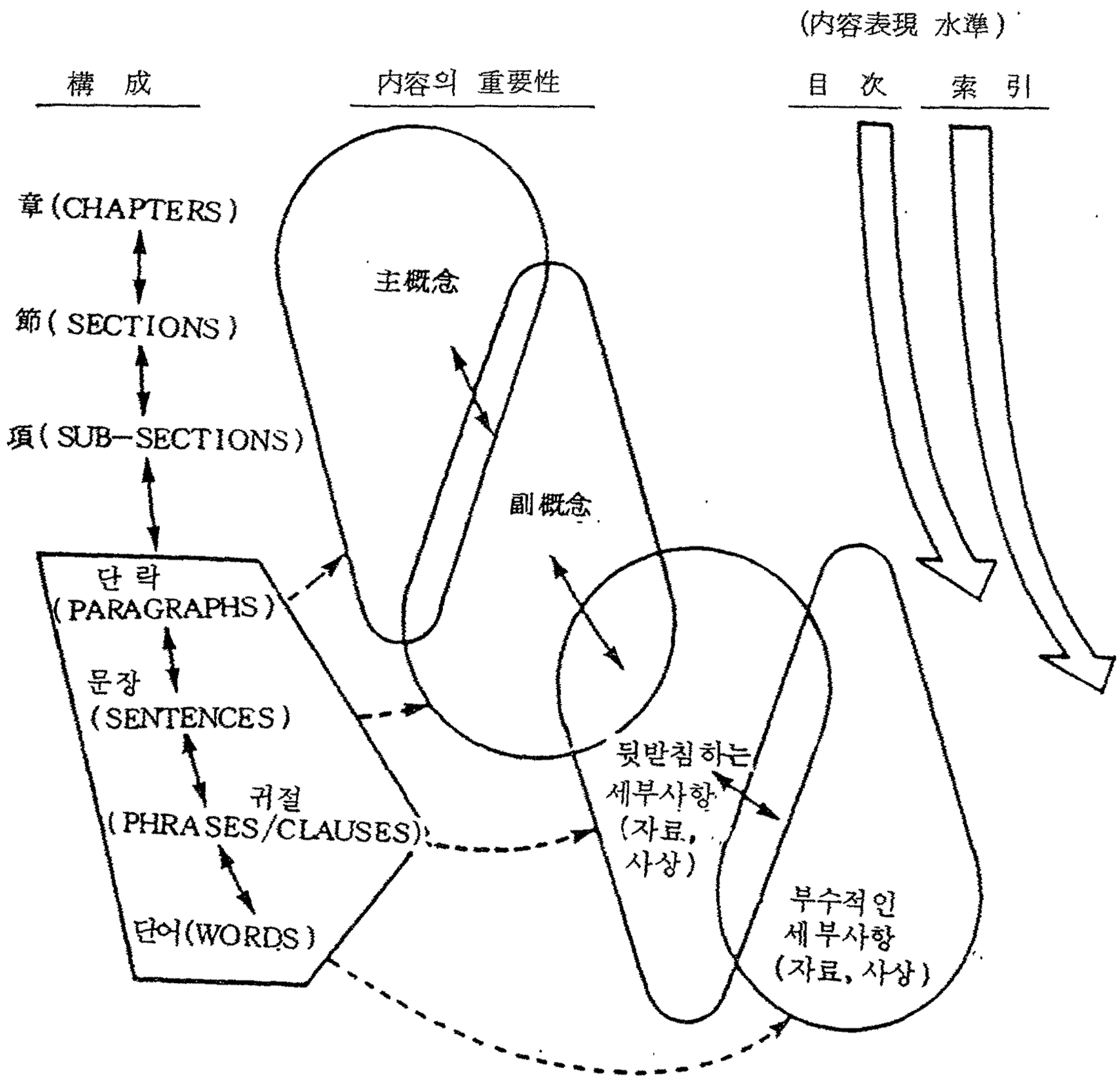


5. 索引 (Book End Index) 의 價値

Bernstein 등에 의하면 單行本의 구성과 주제와 관련된 내용은 〈圖 3 〉과 같이 나타낼 수 있다.[21]

시소러스 作成에서 單行本 全體를 통해서 필요한 用語를 수집하기 어려우므로, 전체 내용을 일별할 수 있는 부분인 目次와 索引을 검토의 대상으로 한다. 目次와 索引에는 資料의 主槪念과 대부분의 副槪念, 그리고 일부 細部事項의 情報가 담겨져 있으며, 索引에 좀 더 구체적인 情報가 있어 索引語를 중심으로 用語를 수집한다.

〈圖 3〉 單行本에서 情報內容의 表現水準



Ⅲ. 시소러스 作成方法

1. 用語蒐集 및 用語間 關係 定義

分野別 專門 用語辭典과 單行本의 Book End Index를 이용하여 用語와 用語間의 관계를 정의한 후, 수집된 데이터로부터 用語間의 關係를 추가하여 목적한 시소러스를 작성한다.

(1) 用語辭典 利用

用語辭典은 해당 분야의 專門家들이 分野의 主要 用語에 대해 설명한 자료이므

로, 시소러스 作成法의 연역적 기법과 유사하며, 일반적으로 情報利用者가 접근할 수 있는 정도의 전조합 수준 (precoordination level) 이므로 用語의 분할에 대해 별도로 고려할 필요가 없다. 그러나, 專門用語가 아닌 기초분야의 용어는 탈락할 우려가 있다.

用語辭典으로부터 표제어와 표제어를 설명하는 說明文의 주요 용어(說明語) 를 〈圖 4 〉와 같은 構造의 레코드로 입력한다. 하나의 표제어에 대해 여러개의 레코드가 만들어진다.

〈圖 4 〉 入力파일構造(用語辭典)

Rec.No.	表題語 (h_i)	說明語 (d_j)	關係 (R)

(2) 索引 (Book End Index) 利用

索引된 用語가 설명되어 있는 부분(文章 또는 段落)으로부터 索引語와 다른 用語間의 關係를 〈圖 5 〉와 같이 入力한다.

〈圖 5 〉 入力파일構造 (Book End Index)

Rec.No.	索引된 用語 (h_i)	索引語가 있는 文章의 다른 用語 (d_j)	關係 (R)

索引된 用語들은 情報價値가 충분히 있으며, 用語辭典利用의 경우와 같이 情報利用者가 용이하게 접근할 수 있는 전조합 수준이며, 對象資料는 다음과 같다.

① 中·高等學校用 教科書 : 用語辭典에서 수집하기 어려운 基礎用語를 수집할 수 있다.

② 大學教材水準 이상의 單行本 : 用語辭典의 불충분한 분야의 用語蒐集과 用語間의 關係를 정의할 수 있다.

(3) 수집된 用語로부터 用語間의 關係 追加 定義

다음의 方法으로 수집된 用語로부터 用語間(하나의 표제어와 다른 표제어간) 의 關係를 追加定義하여 시소러스의 질을 높인다.

① (1), (2)를 통해 입력되어 있는 파일을 " F "라 하면, " F "는 다음과 같다.

$$F = \{(h_1, d_1, R),\ (h_2, d_2, R),\ \cdots\cdots,\ (h_m, d_k, R)\}$$
$$= \{(h_i, d_j, R) \mid i = 1,2,\cdots\cdots,m, j = 1,2,\cdots\cdots,\ k, R = s,b,n,r\}$$

h_i : 用語辭典의 表題語 또는 單行本의 索引된 用語

d_j : 用語辭典의 說明語 또는 單行本의 索引된 用語와 함께 쓰인 用語

R : h 와 d_j의 關係로 s , b , n , r중 하나의 값을 갖는다.

s : h_i와 d_j의 관계가 동의

b : h_i를 중심으로 d_j가 上位槪念

n : h_i를 중심으로 d_j가 下位槪念

r : h_i와 d_j가 關聯關係

임의 h_i와 h_j 또는 d_k와 d_1, h_i와 d_k간에 관계가 있을 수 있으며 이들 간의 關係를 추가 정의하여야 한다.

② 파일 " F "에서 h_i와 d_j의 위치를 바꾸어 " F_r "이라 하고, F와 "F_r"을 합하여 "F_c"로 한다.

$$F_c = \{h_i, d_j, R\} \cup \{d_j, h_i, R'\}$$
$$= \{H_c, D_c, R \mid c = 1,2,\cdots\cdots q,\ R = s,b,n,r\}$$

R′ : h_i와 d_j가 階層關係이면 값을 바꾸어 준다.

즉, b는 n으로, n은 b로 한다.

이 과정에서 모든 h_i가 D_c로 옮겨졌으므로 D_c간의 관계를 定義하면 h_i와 h_j 또는 d_i와 d_j, h_i와 d_j간의 관계를 모두 정의한 결과가 된다.

③ 用語는 클러스터를 형성하는 성질이 있으므로 동일한 H_c에는 관계가 다른 여러개의 D_c가 있으며, 이들 D_c간에는 어떠한 形態이던 關係가 있을 확률이 높다. 따라서, 다음과 같은 파일을 만들어 동일한 H_c와 관계있는 D_c간의 관계를 定義한다. 임의 H_c를 H_{ci}라 하고, H_{ci}와 관계있는 D_c가 p개 있다고 하면, $D_{ci} = \{D_{ci1}, D_{ci2}, \cdots\cdots, D_{cip}\}$이며, 이들로 구성된 파일 F_d는 다음과 같다.

$$F_d = \{(D_{ci1}, D_{ci2}, R_d), (D_{ci1}, D_{ci3}, R_d), \cdots\cdots, (D_{ci1}, D_{cip}, R_d),$$
$$(D_{ci}, D_{ci3}, R_d), (D_{ci2}, D_{ci4}, R_d), \cdots\cdots, (D_{ci2}, D_{cip}, R_d)$$
$$\vdots$$

$(D_{cip-1}, D_{cip}, R_d)\}$

$R_d = \{ s, b, n, r, x \}$

임의 D_{cip} 간의 關係, 즉 R_d 의 값을 결정한다. R_d 는 s,b,n,r,x 가운데 하나의 값을 갖게 되며, R_d 의 값이 x 일 경우는 D_{cip} 간의 關係가 없음을 의미한다.

④ F_d 로부터 $R_d = x$ 인 경우 즉, D_{cip} 간의 關係가 있는 레코드만으로 ②에서와 동일한 요령으로 F_{dc} 를 만든다. F_{dc} 를 F_c 에 합하여 새로운 파일 F_t 를 만든다.

2. 디스크립터 決定

파일 F_t 의 각 用語의 頻度數에 따라 디스크립터를 결정한다. 시소러스의 規模는 축적한 또는 축적하려는 文獻의 數와 文獻當 평균적으로 부여되는 索引語의 수에 따라 결정하는 Houston 등이 제안한 다음 式을 이용하여 예측할 수 있다.[22]

$T = 3{,}300 \log(P + 10{,}000) - 12{,}600$

$P = D \times d$

T : 시소러스에 收錄한 用語數

D : 蓄積할 文獻數

d : 文獻當 平均 索引語數

예를 들어 $D = 100{,}000$ 이고, $d = 10$ 이라면 $T = 7{,}214$ 가 된다.

이와 같은 方法으로 T가 결정되면, F_t 로부터 頻度數에 따라 분포도를 작성한다. 분포도에서 頻度數가 높은 부분과 낮은 부분을 제외한 중간 부분의 用語로 T를 만족시키는 範圍의 用語를 디스크립터로 한다. 用語는 비디스크립터로 한다.

用語 頻度數에 따른 用語數가 標準定規分布를 보이며, 頻度數가 높은 용어는 頻度數가 낮은 用語와 같은 정도로 디스크립터로 선택할 가치가 없다.[23] 고 가정하면 다음과 같은 方法으로 선정 범위를 결정할 수 있다.[24]

평균이 μ 이고 標準偏差가 σ 인 정규분포의 확률밀도 함수는,

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty \text{이다.}$$

수집된 用語의 用語數와 頻度數가 〈圖 6 〉과 같다면, $T=\int_{l}^{h} f(x)dx$를 만족시키는 l과 h를 구하여, 頻度가 l 이하이거나 h 이상인 용어들을 비디스크립터로 한다.

$$\int f(x)dx = \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\chi-\mu)^2}{2\sigma^2}} dx \text{ 에서 } \frac{x-\mu}{\sigma} = z \text{ 라 하면}$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \text{ 가 된다.}$$

수집된 用語의 총수가 T_t 이고, 결정된 디스크립터의 수가 T라 하면 표준정규분포표를 이용하여, $\int_{l_z}^{h_z} f(z)dz = \frac{T}{T_t}$를 만족시키는 l_z와 h_z를 구한다.

$$1 - \frac{T}{T_t} = \int_{-\infty}^{l_z} f(z)dz + (1 - \int_{-\infty}^{h_z} f(z)dz) \text{ 이며,}$$

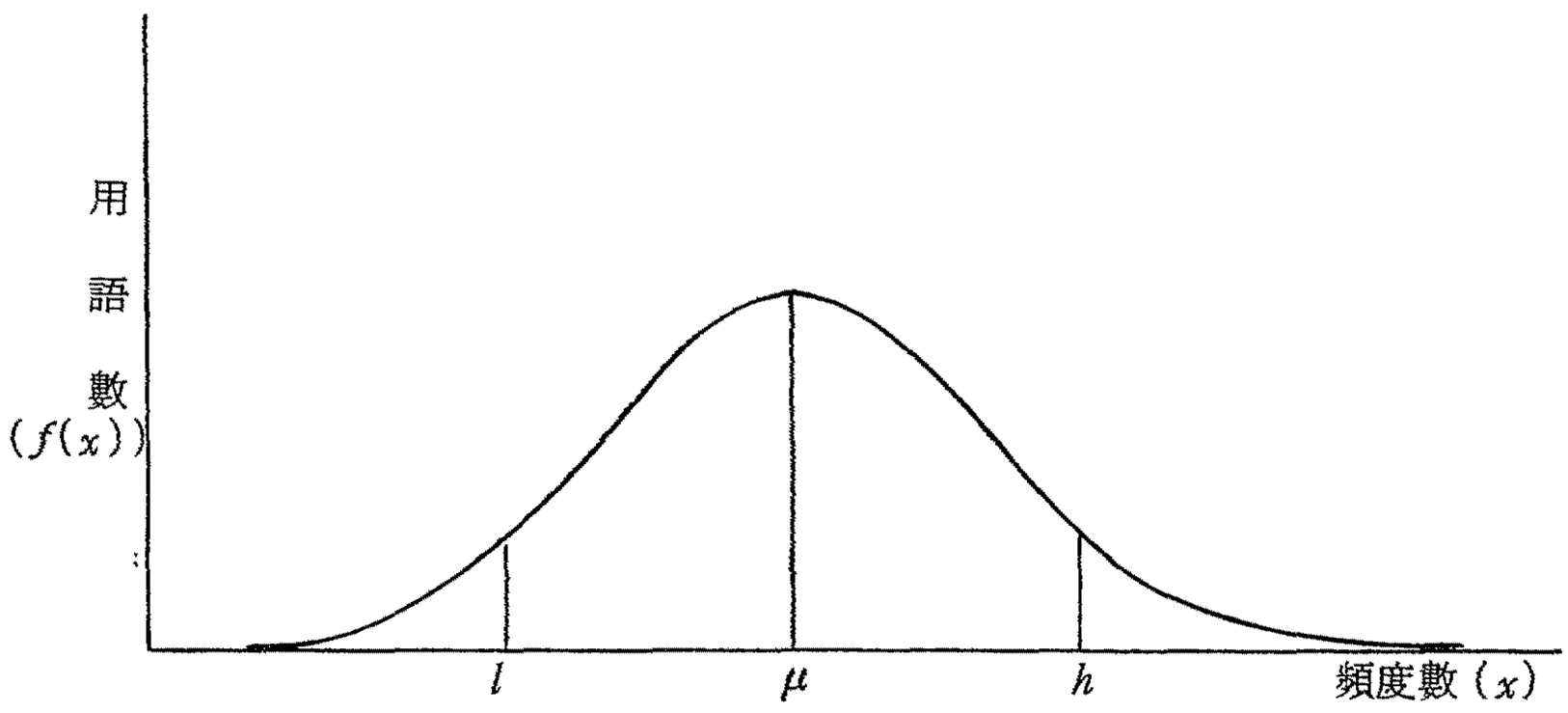
頻度數가 낮은 用語와 頻度數가 높은 용어는 디스크립터로 선택할 가치가 같으므로 $\int_{-\infty}^{l_z} f(z)dz = 1 - \int_{-\infty}^{h_z} f(z)dz$ 이다.

여기서, $z=\frac{x-\mu}{\sigma}$, $\sigma=1$ 이므로, l과 h는 각각 $l=\mu+l_z$, $h=\mu+h_z$ 이다.

예를 들어 $T_t=10{,}000$ 이고, $T=7{,}214$ 平均이 4 라면, 표준정규분포표로부터 $l_z=-1.08$ 이 된다.

마찬가지로 $\int_{-\infty}^{h_z} f(z)dz = 1 - \int_{-\infty}^{l_z} f(z)dz = 0.8607$, $h_z=1.08$, $l=4-1.08=$

〈圖 6〉 用語數와 頻度數



$= 2.92$ 이며, $h = 4 + 1.08 = 5.08$ 이다. 따라서, 頻度가 3 − 5 인 用語를 디스크립터로 한다.

3. 비디스크립터의 處理

결정된 디스크립터들은 시소러스形態로 出力하고, 비디스크립터는 다음 방법으로 디스크립터와 연결시켜 참조토록 한다.

① 同義語가 있는 경우는 同義語를 참조하도록 "USE 同義語"로 표시한다.

② 상위어가 있는 경우는 비디스크립터와 관계가 있는 用語들 가운데 同義語가 없고, 디스크립터로 상위어가 있는 경우에는 상위어를 참조하도록 "USE 상위어"로 表示한다. 상위어가 여러개 있을 때에는 모든 상위어를 모두 참조하도록 한다.

③ 하위어가 있는 경우는 同義語와 上位語가 없고, 디스크립터로 下位語만 있는 경우는 "USE 下位語"로 표시한다. 하위어가 여러개 있을 때에는 모든 하위어를 모두 참조하도록 한다.

④ 關聯語만 있는 경우는 시소러스의 收錄對象에서 제외한다.

4. 시소러스의 出力形態

앞의 과정을 통하여 用語間의 관계가 정의되어 있는 파일을 원하는 형태로 출력하여 최종적인 시소러스가 만들어진다. 일반적으로 시소러스의 出力形態는 가나다(알파벳)순 排列(alphabetical display), 계층 또는 體系的 排列(systematic display), 그래픽 排列(graphic display)로 나누어지며,[25] 다음은 다양한 形態로 작성된 시소러스의 例이다.

(1) 가나다(알파벳)順 排列

모든 수집한 用語를 가나다順으로 배열하고 이와 관계가 있는 語彙를 배열된 각 用語 다음에 배열한 가장 전형적인 시소러스의 出力形態가 〈圖 7〉이다.

用語間의 關係는 표시하지 않고 비디스크립터는 수록하지 않으며 전거 리스트(authority list)와 같은 形態를 취한다(〈圖 8〉 參照).

〈圖 7〉 가나다(알파벳)순 排列의 例[26]

automatic telephone systems
BT telephone systems
TT telecommunication systems
RT electronic switching systems
telephone equipment
telephony
CC B6210D C3370C
DI January 1973

automatic teller machines
BT EFTS
TT computer applications
RT bank data processing
banking
point of sale systems
DI January 1985
PT electronic funds transfer systems
point of sale systems

automobile industry
UF motor industry
BT industries
TT industries
RT automobiles
DI January 1973

automobiles
UF automobile electronics
cars (vehicles)
BT road vehicles
TT vehicles
RT automobile industry
road traffic
CC B8520 B8520B B8620 C3360B C3350Z
DI January 1973

〈圖 8〉 主題分野別 디스크립터 가나다(알파벳)순 排列의 例[27]

450 READING

ADULT LITERACY
ADULT READING PROGRAMS
BASAL READING
BASIC READING (1967 1980)
BEGINNING READING
CLOZE PROCEDURE
CONTENT AREA READING
CONTEXT CLUES
CORRECTIVE READING
CREATIVE READING (1966 1980)
CRITICAL READING
DECODING (READING)
DEVELOPMENTAL READING (1966 1980)
DIRECTED READING ACTIVITY
EARLY READING
ELECTIVE READING (1966 1980)
EYE VOICE SPAN
FACTUAL READING (1966 1980)
FUNCTIONAL LITERACY
FUNCTIONAL READING
GROUP READING (1966 1980)

470 PHYSICAL EDUCATION AND RECREATION

ARCHERY
ATHLETES
ATHLETIC COACHES
ATHLETIC FIELDS
ATHLETICS
BASEBALL
BASKETBALL
BICYCLING
CALISTHENICS
CAMPING
CHILDRENS GAMES
COMMUNITY RECREATION PROGRAMS
DAY CAMP PROGRAMS
EXERCISE
EXERCISE (PHYSIOLOGY) (1969 1980)
EXTRACURRICULAR ACTIVITIES
EXTRAMURAL ATHLETICS
FIELD HOCKEY
FOOTBALL
GAMES

(2) 階層的 排列

가나다順으로 用語를 배열하되 모든 下位槪念語를 계층적으로 배열한다. 가나다순 排列(〈圖 7〉)에서는 下位槪念語의 차하위개념어는 나타나지 않는다(〈圖 9〉 參照).

用語의 가나다順은 고려하지 않고 分類體系에 따라 用語를 배열하는 形態이다(〈圖 10〉 參照).

〈圖 9〉 가나다(알파벳)순 階層排列의 例[28]

DOMESTIC TRADE
uf *trade, domestic*
BT1 *trade*

DOMESTICATED BIRDS
BT1 birds
NT1 pigeons
NT1 poultry
NT2 capons
NT2 chickens
NT3 bantams
NT3 broilers
NT3 chicks
NT4 day old chicks
NT3 cockerels
NT3 pullets
NT2 drakes
NT2 ducks
NT3 ducklings
NT3 wild ducks
NT2 geese
NT3 ganders
NT3 goslings
NT2 guinea fowl
NT2 hens
NT2 turkeys
NT3 poults
NT3 toms

DON
BT1 goat breeds
BT2 breeds
BT1 horse breeds

DONAX
BT1 mollusca

DONGOLA
BT1 horse breeds
BT2 breeds

doob
USE cynodon dactylon

DOOR TO DOOR SALES
BT1 marketing techniques
BT2 marketing
BT2 techniques
BT1 retail marketing
BT2 marketing channels
BT3 marketing

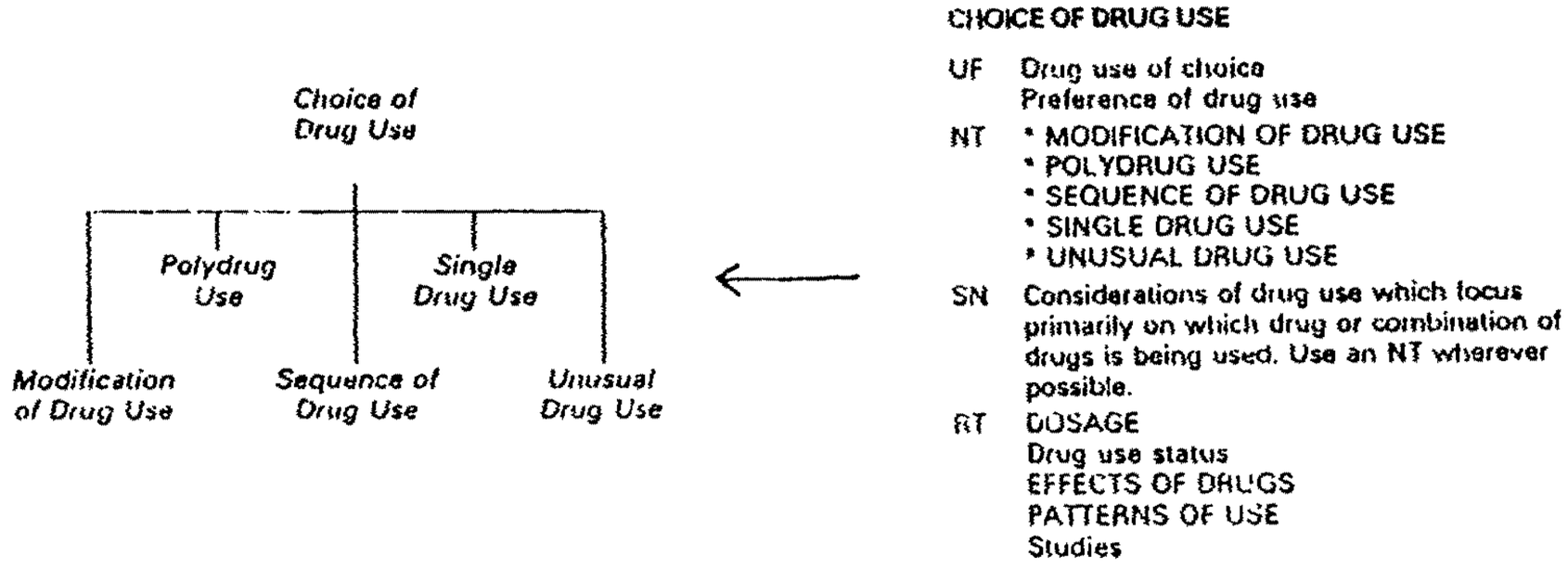
DOORS
BT1 buildings
rt gates

〈圖 10〉 分類體系에 따른 階層的 排列의 例[29]

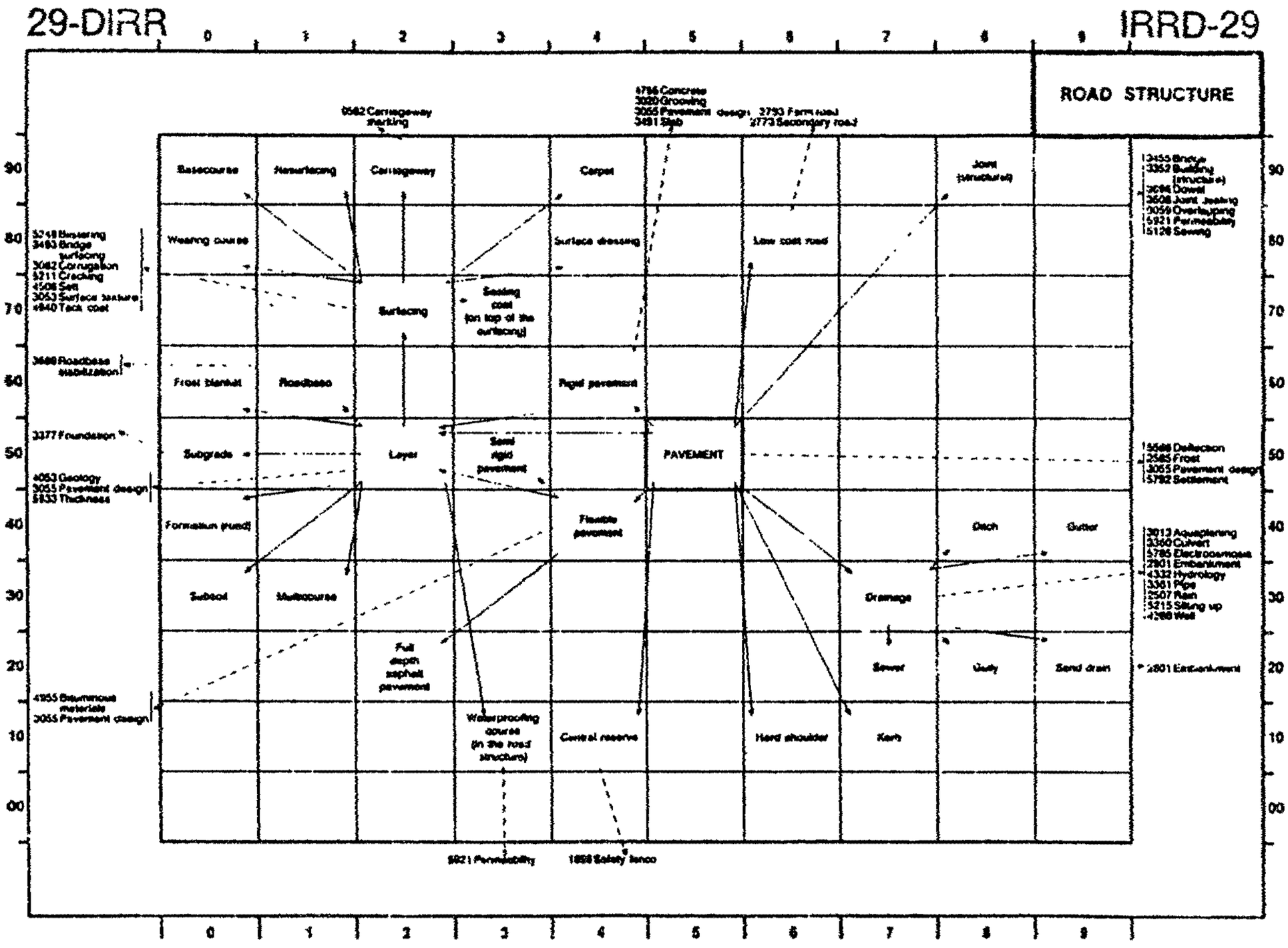
K	**Electrotechnology**
KB/KO	**Electrical engineering** (continued)
KE/KJ	**Electrical equipment** (continued)
KIP	**Electrical protection equipment** (continued)
KIP.V	Electric contact protection *—Electric contacts KNR
	(By construction)
KIP.W	Double electrical insulation *—Electrical insulation CYB.K *—Electrical insulation devices KNX
	(By connection to earth)
KIP.X	Earthing =Earth (electric) =Earthing systems =Electric grounding =Grounding (electric) *—Earthing reactors KHC.E
KIP.XE	Earth electrodes *—Electrodes KNW
KIP.XH	Earth conductors =Protective conductors *<Electric conductors KNN
KIP.XN	Earthing switches =Automatic earthing switches *<Switches KJH
KIP.XR	Neutral conductors *<Electric conductors KNN
KJ	**Switchgear** *>Fuses KIP.M *—Bus-bars KNN.B *—Electric control equipment KIB *—Switching substations KDS.SH
KJC	Circuit-breakers =Air-break circuit-breakers =Air circuit-breakers *>Earth-leakage circuit-breakers *>Relay circuit-breakers KIP.PC *—Operating time MBC.DP *—Switch-fuses KJH.C *—Switches KJH
	(By size)
KJC.C	Miniature circuit-breakers *—Fuses KIP.M
	(By operating medium)
KJC.E	Oil circuit-breakers
KJC.G	Gas-blast circuit-breakers
KJC.GC	Air-blast circuit-breakers
KJC.H	Vacuum circuit-breakers *<Vacuum devices NPT
	(By design)
KJC.M	Tri-pole circuit-breakers =Triple-pole circuit-breakers

(3) 그래픽 排列

〈圖 11〉 나무 構造 出力의 例[30]



〈圖 12〉 用語맵 (Term Map)의 例[31]



特定主題分野를 수로 표시하는 표의 형태이며, 廣義語가 中央에 위치하고 下位概念語 등은 화살표로 표시하며, 화살표를 사용하므로 arrow graph라고 한다.

Ⅳ. 實驗 및 結果分析

1. 實 驗 對 象

實驗對象은 이 實驗에 객관성을 부여하기 위해 일반적으로 서점에서 구입할 수 있는 用語辭典을 사용하였다.

金屬用語辭典에서 [32] 金屬材料分野의 表題語 339個를 선정하고, 선정된 표제어를 설명하는 說明語 1,037個를 〈圖 13〉과 같은 레코드構造를 갖도록 파일 F에 1,521個의 레코드를 입력하였다.

〈圖 13〉 實驗에서 사용한 入力 파일의 構造

REC NO(4BYTE)	表題語 (30BYTE)	說明語(30BYTE)	關係(1BYTE)

2. 實 驗 方 法

入力된 파일로부터 表題語間 또는 說明語間의 관계를 추가 정의하기 위해 최초 入力 파일F로 Fc에서 임의의 "用語A"와 "關係R"의 關係가 있는 用語 B와 C가 있을 때, B와 C의 관계를 定義하였다.

이때 〈圖 14〉와 같은 파일 Fd를 만들어 A와 B 및 A와 C의 關係形態에 따라 B와 C가 관계가 있는 정도는 얼마나 되며, B와 C의 關係形態는 어떠한 성향을 보이는가를 〈表 3〉, 〈表 4〉로 만들어 분석하였다.

여기서, R은 s, b, n, r, x 가운데 하나의 값을 가지며, K는 ss, rr, bb, nn, sb, sr, sn, rn, rb, bn 가운데 하나의 값을 갖는다. 이들의 의미는 다음과 같다.

〈圖 14〉 關係 追加 定義 파일(Fd)의 構造

REC NO (4BYTE)	用 語 B (31BYTE)	用 語 C (30BYTE)	關 係 R (1BYTE)	K(2BYTE)

ss : B와A의 관계가 s이고, C와A의 관계가 s일 경우,
rr : B와A의 관계가 r이고, C와A의 관계가 r일 경우,
bb : B와A의 관계가 b이고, C와A의 관계가 b일 경우,
nn : B와A의 관계가 n이고, C와A의 관계가 n일 경우,
sb : B와A의 관계가 s이고, C와A의 관계가 b이거나,
C와A의 관계가 s이고, B와A의 관계가 b일 경우,
sr : B와A의 관계가 s이고, C와A의 관계가 r이거나,
C와A의 관계가 s이고, B와A의 관계가 r일 경우,
sn : B와A의 관계가 s이고, C와A의 관계가 n이거나,
C와A의 관계가 s이고, B와A의 관계가 n일 경우,
rn : B와A의 관계가 r이고, C와A의 관계가 n이거나,
C와A의 관계가 r이고, B와A의 관계가 n일 경우,
rb : B와A의 관계가 r이고, C와A의 관계가 b이거나,
C와A의 관계가 r이고, B와A의 관계가 b일 경우,
bn : B와A의 관계가 b이고, C와A의 관계가 n이거나,
C와A의 관계가 b이고, B와A의 관계가 n일 경우,

3. 實驗結果 및 分析

(1) 實 驗 結 果

作成된 파일 Fd의 크기는 6,186 레코드이며, 최초 入力 파일 F와 157레코드가 중복되었다. Fd에서 관계가 있는 것으로 판명된 레코드는 1,126 개로 중복되는 부분을 제외하면 969 레코드이었다.

Fd파일에서 관계있는 레코드들의 성향을 分析한 結果를 〈表 3〉으로, Fd에서 F와 중복부분을 제외한 순수하게 새로 定義된 부분만의 內容을 〈表 4〉로 나나내 보았다.

(2) 實驗結果分析

〈表 3〉과 〈表 4〉에서 대체적인 성향에는 변동이 없다. 따라서 重複性 여부는 아무런 영향을 미치지 않음을 알 수 있으며, 用語間의 친화성 및 用語間 관계의 성향을 정확히 파악하는데는 중복분이 〈表 3〉이 유리할 것이다. 〈表 3〉을 分析해 본 결과 다음과 같은 현상을 발견하였다.

〈表 3〉 用語間 關係의 性向

K	FD	FC	FR	S	N-B	R	S/FD	S/FC	N/FD	N/FC	R/FD	R/FC
ss	27	23	.85	12	5	6	.44	.52	.19	.22	.22	.26
rr	2,638	329	.12	14	48	267	.01	.04	.02	.15	.10	.81
bb	66	36	.55	1	25	10	.02	.28	.38	.69	.15	.28
nn	663	113	.17	7	19	87	.01	.06	.03	.17	.13	.77
sr	386	125	.32	1	4	120	.00	.01	.01	.03	.31	.96
sb	74	46	.62	1	37	8	.01	.02	.50	.08	.11	.17
sn	94	40	.43	0	31	9	.00	.00	.33	.78	.10	.23
rb	801	168	.21	2	47	119	.00	.01	.06	.23	.15	.71
rn	1,194	150	.23	2	17	131	.00	.01	.01	.11	.11	.87
bn	243	96	.40	1	73	22	.00	.01	.30	.76	.09	.23
計	6,186	1,126	.18	41	306	779	.01	.04	.05	.27	.13	.69

註 : FD=파일 Fd의 크기, FC=파일 Fd중 관계있는 用語, FR=FC/FD, S=s關係인 用語, N−B=n 또는 b關係인 用語, R=r關係인 用語, S/FD = S/FD, S/FC = S/FC, N/FD = N−B /FD, N/FC = N−B/FC.

〈表 4〉 用語間 關係의 性向(重複分 除外)

K	FD*	FC*	FR*	S*	N-B*	R*	S/D*	S/C*	N/D*	N/C*	R/D*	R/C*
ss	22	18	.85	7	5	6	.32	.39	.22	.28	.27	.33
rr	2,579	270	.12	10	30	230	.00	.04	.12	.11	.09	.85
bb	63	33	.55	1	24	8	.02	.30	.38	.72	.13	.24
nn	653	103	.17	4	18	81	.01	.04	.03	.17	.12	.79
sr	373	112	.32	1	2	109	.00	.01	.01	.02	.29	.97
sb	66	38	.62	1	31	8	.02	.03	.47	.82	.12	.21
sn	92	38	.43	0	30	8	.00	.00	.33	.79	.09	.21
rb	767	134	.21	1	33	100	.00	.01	.04	.25	.13	.75
rn	1,174	130	.23	1	16	113	.00	.01	.01	.12	.10	.87
bn	240	93	.40	0	71	22	.00	.00	.30	.76	.09	.24
計	6,029	969	.18	26	260	683	.00	.03	.04	.27	.11	.70

註 : FD*=파일 Fd의 크기(重複 除外), FC*=파일 Fd중 관계있는 用語(重複 除外), FR*=FC/FD(重複 除外), S*=s 관계인 用語(重複 除外), N-B*=n 또는 b관계인 用語(重複 除外), R*=r관계인 用語(重複 除外), S/D*=S*/FD*, S/C*=S*/FC*, N/D*=N-B*/FD*, N/C*=N-B*/FC*

① 用語 A와 동의(s) 관계가 있는 두 用語 B와 C는 친화성이 대단히 크다(K = ss). 用語 B와 C가 관계가 있을 確率은 85%이며, 平均은 18%이다.

② 用語 A와 關聯(r) 관계가 있는 用語 B와 C는 친화성이 작다(K = rr). 平均보다 낮은 12%를 보였다.

③ 用語 B와 C 가운데 하나만 관계가 있을 때 비교적 친화성이 높은 경향을 보이고 있다(K = sb, K = sn). 다른 한 用語가 상하(n-b) 관계일 때 62% 및 43%, 관련관계일 때 32%를 각각 보였다.

④ 두 用語 B, C가 A보다 上位槪念語일 때에도 비교적 친화성이 높음을 보였다(K = bb). 55%가 관련있는 것으로 판명되었다.

⑤ 관계가 있는 용어간의 성향은 A와 B, C간의 관계가 유사함을 보이고 있다. A와 B, C간의 관계가 關聯(r) 關係이면 대부분 關聯關係를 가지며(rr : 81%, sr : 96%, rb : 71%, rn : 87%), 上下關係인 경우는 上下(n-b) 關係를 갖는 성향을 보였다(sb : 80%, sn : 78%, bb : 69%, bn : 76%). 그러나 下位槪念語間에는 關聯關係가 많은 比重(77%)을 차지하고 있다. 시소러스를 작성하는 것은 복잡하고 어려운 일이므로, 처음부터 완벽한 시소러스를 만들기는 사실상 곤란하다.[33] 가능한 範圍內에서 시소러스를 作成하고, 使用하면서 점차 改正 補完하는 方法이 效果的이다.

따라서, 뒤의 현상에서 친화성이 큰 부분에 대하여만 用語間의 관계를 追加 定義하는 方法이 效率的이다. 즉, 用語 A와 관계가 있는 두 용어 B, C간의 관계는 용어 A와의 관계가 同義 또는 類似關係인 경우(ss, sb, sn, sr)와 上位槪念語인 경우(bb)만 用語間의 관계를 推定 定義하는 方法이 經濟的이다.

V. 結　　論

이 硏究를 통해 다음과 같은 점들을 발견할 수 있었다.

첫째, 시소러스 作成時 用語의 蒐集에서 문제가 되는 용어의 전조합 수준은 用語辭典의 表題語 및 單行本의 索引語가 일반적으로 특정 주제에 접근하려 할 때 찾아보는 수준이므로 적합하며, 分野別 用語辭典과 單行本의 索引을 사용함으로써 특정 기사로부터 用語를 수집하였을 때 발생할 수 있는 주제의 偏重性을 배제할 수 있다.

둘째, 統計的 方法을 도입하여, 처리해야 할 文獻數에 따른 디스크립터의 수 즉, 시소러스의 規模를 실증적으로 정하였다.

셋째, 實驗을 통하여, 일차적으로 定義된 用語間의 관계를 분석하고, 용어간의 관계에 따른 用語間의 친화성 및 성향을 밝혔다. 즉 용어A와 관계가 있는 두 용어 B, C간의 관계는 用語 A와의 관계가 同義 또는 類似關係인 경우(ss, sb, sn , sr)와 上位槪念語인 경우(bb)만 용어간의 관계를 추가 정의하는 方法이 經濟的이며, 이를 토대로 效率的으로 用語間의 관계를 추가 정의함으로써 시소러스의 내실을 기할 수 있다.

시소러스의 단점은 統制語의 장단점에서 거론한 바와 같이 작성에 많은 費用과 시간이 소요되기 때문에 시소러스의 利點을 충분히 알고 있으면서도 쉽게 시도하기 어렵다. 그러나 최근 시소러스를 이용한 專門家 시스템의 開發에 관한 硏究結果가 다수 발표되면서 이에 대한 관심이 다시 늘어가고 있는 추세이다. 英國의 電氣工學會(IEE : Institution of Electrical Engineering)에서 제작한 全世界의 電氣, 電子, 物理, 컴퓨터, 制御 및 情報技術 등의 情報를 수록하는 INSPEC 데이터베이스에서 1988년 한해 동안 이와 관련된 論文을 조사한 결과 모두 17件이 收錄되어 있었다. 장차 作成된 시소러스를 이용하여 專門家 시스템을 開發하는 문제도 硏究해 볼만한 과제 가운데 하나이다.

〈參 考 文 獻〉

1. 「産業情報管理」, 産業硏究院, 1987.
2. 사공 철, 「情報檢索에 있어서의 Thesaurus導入에 관한 基礎硏究」, 延世大學校 碩士學位論文, 1974에서 再引用.
3. *Chemical Abstracts Service 1988 Annual Report.*
4. 사공 철, 「情報檢索論」, 아세아문화사, 1984.
5. Rowley, J.E., *Abstracting and Indexing*(2nd ed.), Clive Bingley Ltd., 1988.
6. Rowley, J.E., *ibid.*
7. *ISO 2788*, "Documentation-Guidelines for the Establishment and Development of Monolingual Thesauri," 1974.

8. 岡谷大, "ターミノロジー方法：ドキュメンテーションへの應用とAIの關係," 「情報の科學と技術」, 37 (9), 1987, pp.405 ~ 412.

9. Felber, H., " International Standardization of Terminology : Theoretical and Methodological Aspects," *Intl. J. Soc. Lang.*, 23, 1980, pp.65 ~ 79.

10. Felber, H., *ibid.*

11. 사공 철, 「情報檢索에 있어서의 Thesaurus導入에 관한 基礎研究」, 延世大學校 碩士學位論文, 1974.

12. *Thesaurus of Engineering and Scientific Terms*, DOD, 1967.

13. *NASA Thesaurus*, National Aeronautics and Space Administration, 1988.

14. 「JICST科學技術用語シソーラス」, 日本科學技術情報センター, 1987.

15. *ROOT Thesaurus*, BSI, 1988.

16. Kim, C., " Theoretical Foundation of Thesaurus-construction and Methodological Consideration for Thesaurus-Updating," *Journal of the American Society for Information Science*, March-April, 1973, pp.148 ~ 156.

17. *BS5723*, " British Guide to Establishment and Development of Monolingual Thesauri," British Standard Institution, 1987.

18. *BS5723-1987*.

19. Yu, C. T., " Single-Pass Method for Determining the Semantic Relationships between Terms," *Journal of the American Society for Information Science*, November, 1977, pp.345 ~ 354.

20. Lancaster, F. W., *Vocabulary Control for Information Retrieval* (2nd ed.), Arlington, Virginia, Information Resources Press, 1986, pp.233 ~ 234.

21. Bernstein, L. M., Williamson, r.E., " Testing of a Natural Language Retrieval System for a Full Text Knowledge Base," *Journal of the American Society for Information Science*, 35(4), 1984, pp.235 ~ 247.

22. Houston, N., Wall, E., " The Distribution of Term Usage in Manipulative Indexes," *American Documentation*, April, 1964, pp.105 ~ 114.

23. Salton, G.外, *Introduction to Modern Information Retrieval*, McGraw-Hill Book Co., 1983, pp.59 ~ 63.

24. Butler, C., *Statistics in Linguistics*, Basil Blackwell Ltd., 1985.

25. *BS5723-1987*.

26 . *INSPEC Thesaurus*, Institution of Electrical Engineers, 1987.

27 . *ERIC Thesaurus Descriptors*, Education Resources Information Centre, 1982.

28 . *CAB Thesaurus*, Commonwealth Agriculture Bureau, 1982.

29 . *ROOT Thesaurus*, British Standard Institution, 1985.

30 . *ISDD Thesaurus*, The Institute for the Study of Drug Dependence, 1980.

31 . *IRRD Thesaurus*, OECD, 1985.

32 . 「金屬用語辭典」, 서울 : 성안당, 1988.

33 . 노환주, 「Thesaurus를 이용한 情報檢索시스템 設計에 관한 硏究」, 東國大 碩士學位論文, 1985.