情報檢索에 있어서 Thesaurus의 導入에 관하여

司 空 哲*

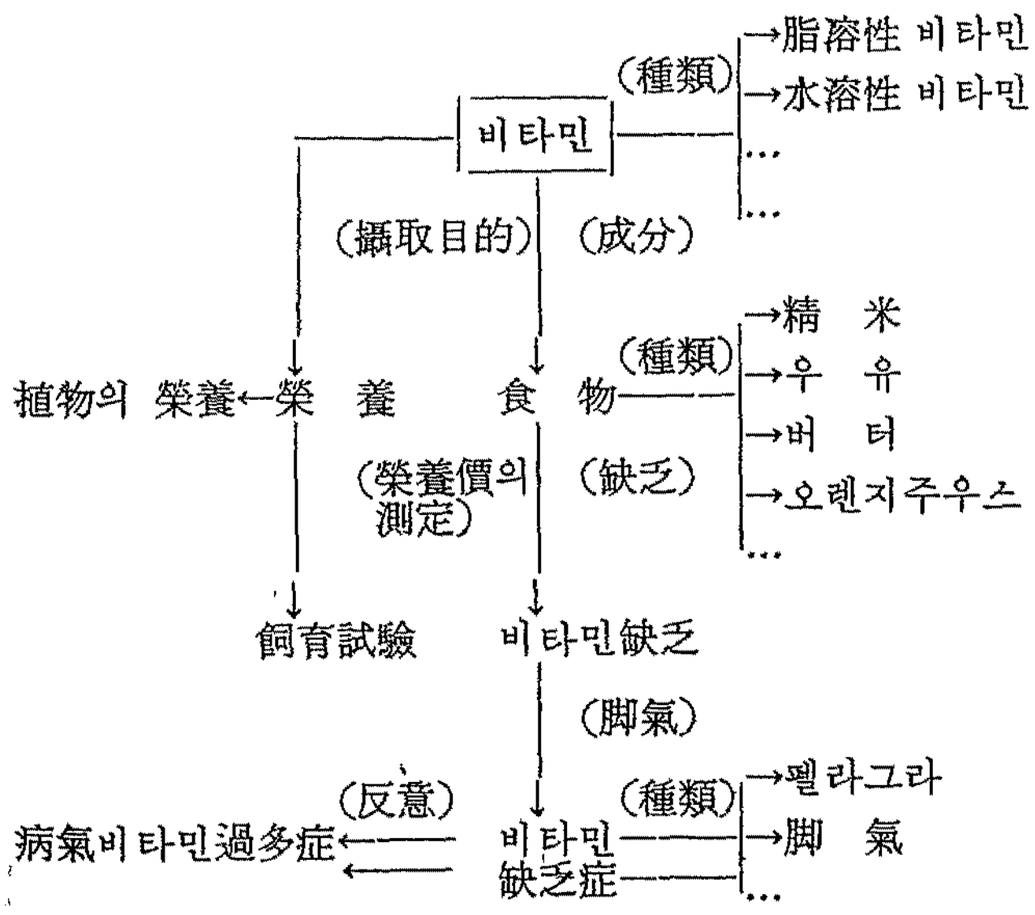
1. 情報檢索에 있어서 Thesaurus의 必要性

情報를 蓄積하기 위하여 索引作成時에 使用되는 言語를 索引語 또는 檢索語라 부른다. 그런데 文獻은 自然語(文獻語)로 쓰여지며, 索引語는 自然語에 의하여 作成되기 때문에 同一한 用語에 대하여도 文獻生産者(著者), 索引作成者, 利用者間에는 서로 學歷, 經驗, 言語習慣 등이 다르기 때문에 여러가지로 表現될 수 있어 檢索에 많은 蹉跌을 가져오게 된다.

뿐만 아니라 어떤 한개의 索引語에 관한 文獻을 찾는다고 하여 그 索引語 하나만으로는 滿足할 만한 調査가 될 수 없으며, 槪念의 大小, 關聯語 등도 폭넓게 檢索하여야 할 것이다.

例를 들어 비타민에 관한 文獻을 찾고자 할 때 비타민 1語만으로는 正確한 檢索이 되지 못하며, 아래의 表 1—1[1]과 같이 槪念을 聯合하여 必要語를 檢索하여야 한다. 그렇기 때문에 모든 索引語에 대한 同義語, 同形異義語, 槪念의 大小 또는 上下關聯語 등을 適切히 調節하여 文獻生産者, 索引作成者, 利用者間에 統一的으로 使用할 수 있는 道具(Toll)가 必要하게 되며 이를 情報檢索用 thesauras라 부른다.

表 1—1. 비타민의 槪念聯合



2. Thesaurus의 語義, 區分 및 効果

2.1 Thesaurus의 語義와 發展

thesaurus란 희랍語의 θησαυρο'ς를 語源으로 하여 辭典, 百科事典과 같은 知識의 寶庫를 意味하는 것으로[2], 이 用語가 著述로서 처음으로 쓰인 것은 1852年 P.M. Roget의 「Thesaurus of English words and phrases」[3]로, 여기에 記述되어 있는 thesaurus의 語義는 "어느 槪念에 대하여 그것을 가장 適切히 表現할 수 있는 標目을 選擇하기 위하여 만들어진 語句의 集大成"이라 말하고 있다[4]. 그렇기 때문에 一般辭典은 用語는 알고 있고 그 뜻을 모를 때 使用하는 것에 비하여 Roget의 thesaurus는 그 反對로 뜻(槪念)은 알고 있으나 그 槪念에 該當하는 用語를 알 수 없을 때 使用하는 것이다.

例를 들면 潛水艦이 물에 잠긴 채로 海氣의 吸排를 하기 위하여 使用하는 가느다란 管을 무엇이라고 하느냐 라는 質問이 있을 때 가느다란 管이란 英語로 tube나 pipe이다. Roget의 thesaurus에서 tube나 pipe項을 調査하면 이들 用語와 關係있는 모든 것을 볼 수 있고 그 중에서 snorkel이라는 答을 얻을 수 있게 된다.

그러나 情報檢索 分野에 있어서의 thesaurus란 上述한 뜻과는 區別되어 使用된다. 情報檢索과 關聯하여 이 用語가 처음으로 쓰인 것은 1957年 英國의 Dorking에서 開催된 情報檢索 分類에 관한 國際會議에서 Brownson[5]이 行한 講演에서 비롯된다.

그는 한 文獻에 나타나는 여러 槪念의 關聯을 보다 整備된 用語로 바꾸어 놓는 것이 情報檢索의 問題이며, 解決法으로서 意味를 相互關聯시키는 thesaurus를 適用하는 것이라 하였다.

그후 이 말은 美國의 專門家들간에 많은 관심사가 되어 왔으며, 그중 이 用語를 널리 紹介한 사람은 H.P. Luhn, J.H. Heald, K.F. Heuman, E. wall 등이다. 이들이 말한 thesaurus의 意味를 要約하면 狹義로는 自然語를 統制語로 變換할 때 도움이 되는 手段이며, 廣義로는 語間의 關係를 表示하는 모든 一元的인 表라 말하고 있다[6].

thesaurus의 定義로 가장 最近의 것은 英國의 CRG (Classification Research Group)이 第 173次 會合에서

* KORSTIC 資料部 次長

定한 것이 있다[7]. 즉 "thesaurus란 後組合索引法(Post Coordinate Indexing)과 關聯하여 情報檢索시스팀에 쓰이는 統制語表로서 이 表에는 槪念的으로 組合된 用語의 關係가 指示되어 있고, 統制語表에는 索引作成에 標準으로 쓰이는 參照가 되어 있다"고 定義하고 있다.

IR用 thesaurus가 어떠한 경로에 의하여 아이디어가 생겨났는지 그에 관해 具體的인 論文을 確認치 못하여 알 수 없으나, 主題名標目表가 圖書整理에 끼친 影響은 누구나가 다 認定하는 事實이다. 따라서 文獻情報管理의 中心的인 資料單位가 되는 論文記事의 整理에도 圖書의 主題名標目表와 같은 것이 必要하게 되어 Dorking 會議 前에 美國에서 생각한 方法이라 思慮된다.

實際로 Gillum[8]도 이와 근사한 表現으로 記述하고 있다. 즉 "傳統的인 主題名 典據目錄은 科學技術 文獻의 機械檢索發達에 따라서 대담하게 改編하였다. 索引作成者나 檢索者가 效果的으로 使用할 수 있는 檢索用語의 새로운 構成法이 생기게 되었는데 그 하나가 thesaurus concept라 불리우는 것이라"말하고 있다. 그러므로 thesaurus는 歷史的으로 主題名標目表를 發展시킨 것이고, 機能的으로는 論文記事의 文獻을 取扱하기 위하여 탄생된 主題名標目表라 할 수 있다.

Dorking會議 2年後인 1959年에 ASTIA는 thesaurus가 機械檢索體係에 必要不可缺한 것이라 結論지우고[9] 刊行한 것이 前述한 바 있는「Thesaurus of ASTIA Descriptors」이다. 그후 1961年에 A.I.Ch.E.[10]가, 1962年에는 ASTIA가 改訂版을 發行하게 되자 thesaurus에 의한 索引方法이 認定받게 되어 1963年 美國大統領科學諮問委員會[11]에서는 技術界에 대하여 各 著者는 論文을 發表할 때 後日에 있을 檢索을 쉽게 하도록 할 責任이 있다고 말하고 thesaurus에 있는 keyword로 된 索引의 附記를 勸告하기에 이르렀고, 그후 thesaurus는 全世界的으로 發展하였다.

2.2 thesaurus의 區分

一般的으로 thesaurus는 文獻語를 한개의 檢索語에 對應시키는 手段으로 쓰이고 있으나, 이와 반대로 한 文獻을 多檢索語에 對應시키는 이른바 分析的 方法 또는 파셋方法을 包含하여 thesaurus를 區分하여 보면 아래와 같다[12].

1) 檢索語만의 것.
2) 多文獻을 한개의 檢索語에 對應시키는 것.
3) 한개의 文獻語를 多檢索語에 對應시키는 것,
4) 文獻語를 檢索語로 하는 것.

처음 것은 主題名標目表가 해당되며, 두번째 것은 大部分의 旣刊 thesaurus이고, 세번째 것은 Western Reserve University의「Semantic Dictionary」가 있으며, 그리고 네번째 것은 Wall의 thesaurus가 있다. 그러나 오늘날 使用되는 것은 거의가 두번째 것에 해당한다.

또한 thesaurus를 形態別로 區分하여 보면 아래와 같다[13].

1) 冊字形 thesaurus

종이에 印刷되어 冊字形으로 된 것으로 가장 一般的인 것이다. 組版하여 印刷하는 것과, 電子計算機에서 print out하는 경우, 그리고 마이크로形으로 된 것을 複寫한 것 등이 있다.

2) 機械 thesaurus

電子計算機의 記憶裝置에 格納되어 있는 것.

3) 마이크로形 thesaurus

마이크로필름, 마이크로피시 등과 같은 마이크로形으로 된 것.

그리고 thesaurur에 收錄되어 있는 檢索語의 分野에 따라 다음과 같이 區分한다.

1) 專門 thesaurus

어느 特定主題만을 取扱한 것.

2) 總合 thesaurus

여러 主題를 綜合的으로 包含하고 있는 것.

2.3 Thesaurus의 效果

thesaurus의 效果에 관하여는 Swanson의 硏究가 先驅的이며, SMART시스팀도 有名한 硏究이다.

A. Swanson의 硏究

電子計算機에 의한 自然語原文을 探索實驗한 것으로 內容을 要約하면 아래와 같다.

蓄積; 核物理學 論文 100件

探索; 다음 3가지 方法을 比較하였다.

1) 主題名標目에 의한 從來의 手動式 檢索

2) 論理和와 論理積을 使用한 機械檢索으로 檢索補助手段을 使用치 않음.

3) 論理和와 論理積을 使用한 機械檢索으로 檢索補助手段을 使用함.

質問; 探索者(7名)×手法(3種)×質問(50) = 약 1,000質問

結果; 有效資料에 대한 探索方法別 結果는 다음과 같다.

方法 1) 38%
2) 68%
3) 86%

探索方法에서 檢索補助手段이라 하여 使用한 것이 thesaurus[15]이며, 이 方法이 86%라는 가장 높은 結果

를 보이고 있다.

B. Smart시스템의 硏究

SMART(Salton Margical Automatic Retriever of Texts) 시스템은 完全自動化 情報檢索 시스템을 硏究, 評價를 目的으로 Salton이 中心이 되어 처음에는 Harvard大學에서 始作하여 후에는 Cornell大學에서 계속되어 왔다.

이 시스템의 實驗結果에 관하여 Lesk[16]가 發表한 것 중 thesaurus의 效果에 關係있는 事項은 다음과 같다.

1) 語處理; 同義語辭書를 使用하는 것은 使用치 않는 경우보다 效率이 높다.

2) 句檢索; 句辭書를 使用하는 方法은 同義語辭書를 使用하는 方法보다 效果面에서 크게 뒤진다.

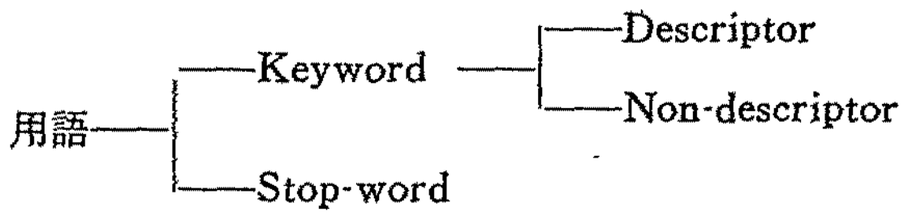
3) 語의 相關; 語의 相關을 使用하는 方法은 同義語辭書에 의한 것보다 效率이 떨어진다.

4) 槪念體係參照; 槪念의 聯合關係를 使用하여 質問檢索語를 展開하는 方法은 使用치 않는 경우보다 效率이 上昇된다.

3. Thesaurus의 構造

3.1 用語의 區分

thesaurus의 構造上으로 用語를 區分하면 다음과 같다[17].



Keyword란 文獻의 內容을 表現하는 重要語를 말하며, 이것은 descriptor와 non-descriptor로 區別된다.

descriptor는 索引作業이나 檢索作業에 索引語(檢索語)로 使用되는 것이고, non-descriptor는 索引語로 使用할 수 없으며, 반드시 descriptor에 參照되는 keyword로, 前者를 standard keyword, 後者를 non-standard keyword라고도 한다. 例를 들어 "電子計算機"와 "電算機"란 keyword가 있을 때 前者를 索引語로 定하면 이것이 descriptor가 되는 것이고 後者는 non-descriptor로 "電子計算機"에 參照가 된다. 그리고 stop-word는 keyword로서 가치가 없는 用語로 "~에 관하여", "硏究"등을 例로 들 수 있다.

3.2 Thesaurus의 構成

Thesaurus의 構成에 관한 標準化에 관하여는 提案[18]이 있을 뿐 確立된 것은 지금까지 없다. 그렇기 때문에 情報檢索에 있어서 Thesaurus의 導入에 관하여 thesaurus의 種類에 따라서 다르게 되어 있다.

그러나 thesaurus의 主體(main body)를 이루고 있는 것은 keyword list이며, 種類에 따라 여러가지로 作成되어 있는 것 즉, 例를 들면 TEST(Thesaurus of Engineering and Scientific Terms)의 경우 Permuted Index 등은 Keyword list의 利用을 돕기 위한 附隨的인 것이기 때문에 Keyword list의 構成을 說明하는 것이 바로 thesaurus의 構成을 意味하는 것이라 하겠다.

thesaurus는 descriptor와 non-descriptor로 構成되며, descriptor와 non-descriptor와의 關係 그리고 descriptor 相互關係로 이루어 진다.

3.3 thesaurus의 項目單位

thesaurus의 項目單位를 記入(entry)이라 부른다. 記入은 記入語와 記入語에 관한 補助事項으로 된다.

記入語란 個個의 keyword를 말하며, 모든 記入은 記入語의 字母順으로 排列된다.

記入語의 補助事項에는 記入語의 主題範圍番號, 限定句, 範疇, 關係語 등으로 되어 있다..

3.4 補助事項

1) 主題範圍番號

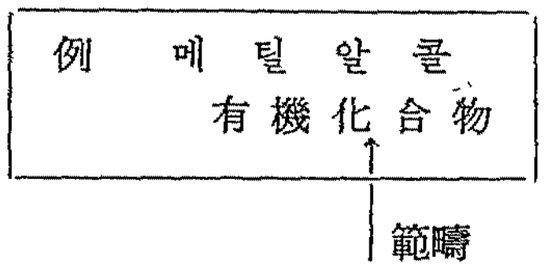
Keyword에 該當하는 主題範圍番號

例	Diene resins	1109

1109라는 번호에서 11은 材料라는 뜻이고 09는 플래스틱을 나타내고 있다.

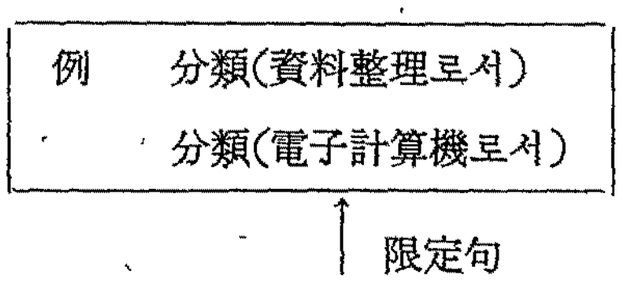
2) 範 疇

Keyword에 부여되어 있는 屬性을 나타낸다.



3) 限定句

同形異義語로 된 keyword의 區別과 意味의 明確을 기하기 위하여 keyword를 修飾하는 句로 keyword의 一部分이 된다. 따라서 이 keyword가 關係語로 쓰일 때에는 限定句도 함께 쓰인다.



4) 關係語

thesaurus가 keyword와 keyword의 關係를 基礎로 하여 構成되는 이상 自然語의 語間에 어떤 關係가 있는

가를 明確히 할 必要가 있다. 물론 그 關係를 詳細히 分析하여 가면 文法論 내지는 言語學의 領域에 이르게 되어 Bernier[19]의 意味語間의 關係, Miller[20]의 用語의 關係, Pery 및 Kent[21]의 分析關係 등을 活用하여야 되지만 情報檢索에 必要한 thesaurus를 理解하기 위한 基本的인 것은 다음 3가지가 있다.

1) 同義語關係
2) 階層語關係
3) 關聯語關係

이들 關係를 表示하는 關係記號에는 아래의 表 3—1에서 보는 바와 같이 thesaurus의 種類에 따라 다르다.

表 3—1. Keyword의 關係記號

關係 \ Thesaurus	COSATI	AICHE	DDC[23]	主題名標目表
同義	USE	SEE	USE	SEE
同義	UF (Used for)	SF (Seen from)	Includes	See
上位	BT (Broader Term)	PO (Post on)	Specific to	See also
下位	NT (Narrower Term)	GT (Generic to)	Generic to	See also
關聯	RT (Related Terms)	RT (Related Terms)	Also see	See also

本稿에서는 가장 많이 採用되는 COSATI[22]方法에 의한다. 위의 表에서와 같이 USE, UF, BT, NT, RT 등 5個의 關係記號가 있다.

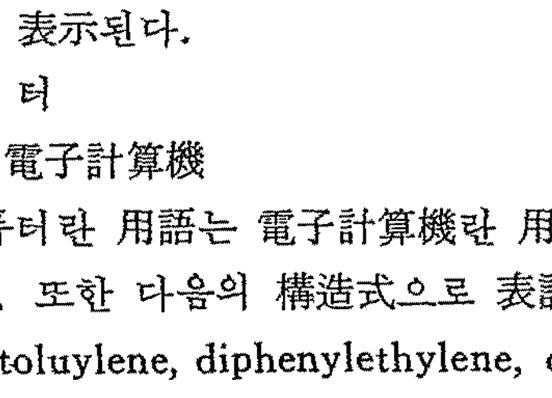
3.5 關係記號

1) USE

同義語 및 準同義語關係에 使用한다. 電子計算機와 컴퓨터란 用語에서 電子計算機를 descriptor로, 컴퓨터를 non-descriptor로 定하였다고 하면 thesaurus에는 아래와 같이 表示된다.

컴 퓨 터
USE 電子計算機

즉, 컴퓨터란 用語는 電子計算機란 用語로 使用하라는 것이다. 또한 다음의 構造式으로 表記한 化合物에는 stylbene, toluylene, diphenylethylene, dibenzal, dibenzylidene 등 5個의 同義語가 있다. 이 構造를 갖고 있는 化合物을 빠짐없이 檢索코자 할 때 5個의 用語 全部를 探索하지 않으면 效果가 없을 것이다. 따라서 同義語가 存在하는 境遇 有效한 檢索結果를 얻기 위하여는 同義語를 整理하고 그중 代表的인 것 하나만을 descriptor로 하고 나머지는 USE로 參照하게 되면 만족스러운 檢索을 할 수 있을 것이다.



2) UF

Used for의 略字로 USE의 逆關係이다.

電子計數機
UF 컴퓨터

즉, 電子計算機란 用語는 컴퓨터란 用語를 대표하여 descriptor로 採用되었다는 것이다.

3) BT와 NT

한 descriptor를 中心으로 한 上位槪念과 下位槪念의 關係를 表示하는 것으로 BT는 broader term의 略字로 上位槪念이고, NT는 narrower term의 略字로 下位槪念을 뜻한다.

電動機란 用語를 例로 들면 電動機를 種類에 따라 細分하면 交直兩用電動機, 直流電動機, 交流電動機로 되고, 다시 直流電動機는 直卷電動機, 分卷電動機, 複卷電動機로 細分된다. 또한 交流電動機도 同期電動機, 誘導電動機 등으로 細分된다. 다시 말하면 電動機는 交直兩用電動機, 直流電動機, 交流電動機를 포함하고, 直流電動機는 直卷電動機, 分卷電動機, 複卷電動機를 포함하고 있다.

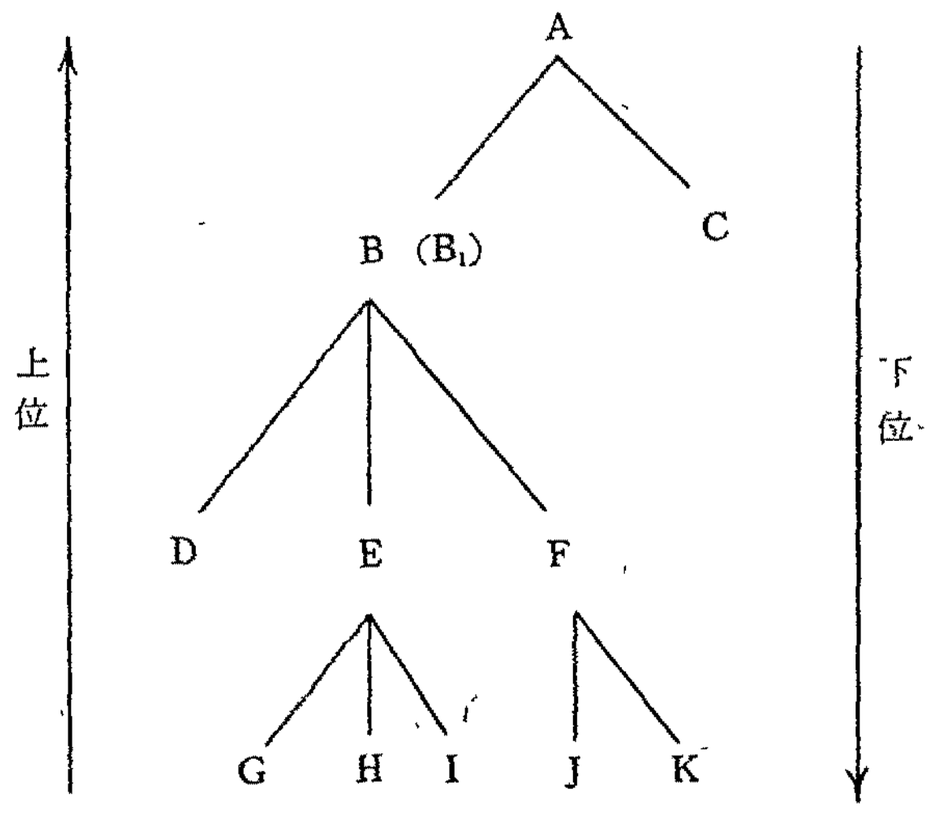
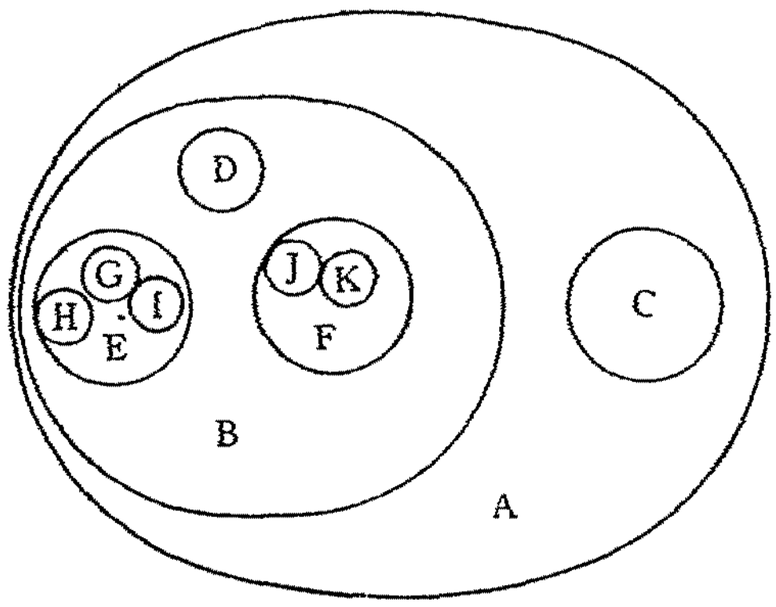


그림 3—1. 用語의 上位語 下位語의 關係

A：電氣機械　　F：交流電動機
B：電動機　　G：直卷電動機
B_1 B의 同義語　　H：分卷電動機
C：發電機　　I：複卷電動機
D：交直兩用電動機　　J：同期電動機
E：直流電動機　　K：誘導電動機

이와 같은 關係에 있을 때 電動機는 交直兩用電動機 등의 上位語이고, 交直兩用電動機 등은 電動機의 下位語가 되는 것이다. 이러한 方式으로 上下의 關係를 定하여 가면 電動機의 上位語는 電氣機械가 되고, 電氣機械의 下位語는 電動機, 發電機가 된다. 이러한 關係를 圖示하면 그림 3—1과 같다.

以上에서와 같이 上下關係가 國際十進分類法등의 階層分類와 같이 階層關係를 나타내고 앞의 그림 3—1과 같은 構造가 될 때 이것을 本構造라 부르고 ABC등을 節이라 하며, 節과 節을 연결하는 線을 枝라고 한다. 그리고 이 關係를 廣義, 狹義의 關係로 圖示하면 아래와 같다.



電動機에 관한 上下關係를 BT, NT로 表示하면 아래와 같다. BT, NT 다음의 數字는 上下의 順位(經路)를 나타내는 것이다.

電動機

BT_1　電氣機械
NT_1　交流電動機
NT_1　交直兩用電動機
NT_1　直流電動機
NT_2　同期電動機
NT_2　複卷電動機
⋮　⋮

4) RT

Related term의 略字로 USE, UF, BT, NT 以外의 關係를 表示하는 것.

電動機와 發電機를 比較하여 보면 前節에서와 같이 電氣機械에 속하나, 兩者사이에는 아무런 階層關係가 存在하지 않는다. 다만 電動機는 動力에 의하여 電流를 일으키게 되는 것이므로 原理的, 機械的으로는 대단히 類似하다. 이와 같은 關係에 있는 用語를 全部 RT로 연결한다.

電動機
RT 發電機

4. Thesaurus의 作成法

thesaurus 作成法에는 收錄할 用語를 各種 收集源에서 引出하여 새로운 것을 作成하는 境遇와 旣刊 thesaurus中 適當한 것을 번역하여 再編成하는 방법 2가지가 있다. 그러나 後者는 硏究할 對象이 되지 못하는 것이기 때문에 本硏究에서는 省略하기로 한다.

thesaurus를 作成하는 作業順序로는 다음과 같이 4段階로 생각할 수 있다.

(1) 用語의 收集
(2) Keyword의 選定
(3) Keyword 相互關係의 決定
(4) 編成

4.1 用語의 收集

用語를 引出하기 위한 收集源으로는 다음과 같이 2種類가 있다. 첫째 統制語가 收錄되어 있는 各種 旣刊 thesaurus, 用語辭典, 主題名標目表, 分類表 등이 있고 다음으로는 文章語가 收錄되어 있는 資料 즉, 各種 學術雜誌, 圖書, 리포트, 便覽 등이 있다. 이중 가장 많이 使用되는 것이 旣刊 thesaurus이며, 其他 主題名標目表를 分析하여 作成하는 境遇, 分類表 또는 分類表의 相關索引을 分析하는 것, 그리고 文章語에 의하는 例를 볼 수 있다.

例를 들면 前述한 바 있는 ASTIA의 것은 主題名標目表에 의한 것이고, AIche의 것은 文章語에서 用語를 抽出하였으며, 分類表에 의한 것은「金屬工學シソーラス」가 있다.

a. 收集規模

用語의 收集源을 各種 thesaurus, 用語辭典 등 統制語로 定할 때에는 이미 調節되어 있는 것이기 때문에 별로 問題가 없으나, 文章語를 收集源으로 할 경우에는 어느 程度의 語數를 取하느냐가 問題인 것이다.

Schwartz[24]에 의하면 專門分野의 用語는 文章에서 대개 4萬語를 抽出하였을 때 약 2,500의 keyword로 定하여 진다고 指摘하고 있으며, Long[25]도 放射線 醫學分野에서 거의 같은 結果를 報告하고 있다. 즉, Long에 의하면 文章에서 약 4～5萬語를 抽出하였을 때 Keyword는 약 2,500語로 安定되며, 그 以後의 增加率은 그림 4—1에서와 같이(점선표시) 매우 낮은 것이다.

結論的으로 어느 專門分野의 Keyword는 文章에서 약

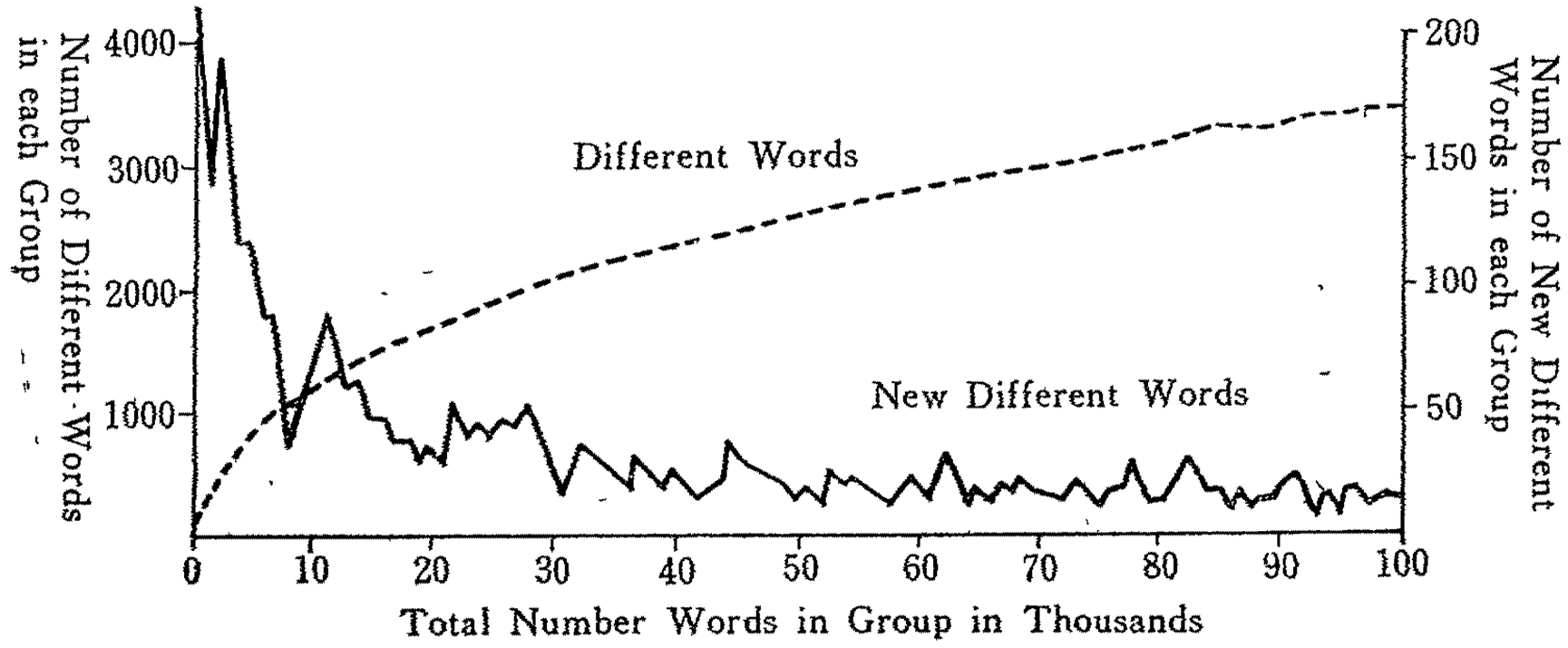


그림 4—1. 專門分野의 新語增加

4萬語가 抽出되었을 때 2,500語로 定하여 지며, 理論的으로 新語의 導入率은 0에 가까운 것이다. Long에 의하면 10萬語에 있어서도 新語의 導入率은 여전하다고 報告하고 있다[26].

單一專門分野가 아니라 總合 thesaurus의 경우에는 TEST[27]의 例와 같이 약 15萬語(一部 社會科學 包含)를 抽出하여 약 2萬語를 Keyword로 定하고 있다.

b. 用語의 單位

用語의 單位를 定하기 위한 規定을 마련하여야 한다. 그렇지 않으면 混雜하게 된다. 例를 들어 "美國特許局의 非金屬化合物의 機械檢索"이란 標題에서 사람에 따라 아래와 같이 用語의 單位를 定하였다고 하면 再現性이 稀薄하게 될 것이다.

金氏：特許局 機械檢索

李氏：美國特許局 非金屬化合物 機械檢索

朴氏：美國特許局 非金屬化合物 機械檢索

單位의 規定은 用語의 客觀性과 再現性을 위하여 必要한 것이다. 規定은 精密한 것이 要求되나, 作業을 簡單히 할 수 있는 方法도 考慮하여야 한다.

이 規定은 作成코자 하는 thesaurus의 目的, 範圍 등에 따라 融通性이 있는 것이어서 한가지로 固定化할 수는 없는 것이다. 다만, 一般的으로 다음과 같은 內容이 取扱된다.

1) 用語의 品詞

用語는 名詞形으로 한다. 英語의 경우 名詞形 以外의 品詞가 必要할 때에는 形容詞 혹은 이와 類似한 것으로 하고, 動詞는 取하지 않는다.

(例) rough는 roughness

catalize는 catalysis

2) 用語의 길이

抽出하는 用語는 連續形으로 取한다. 단, 다음과 같은 경우는 例外로 한다.

1) Keyword 以外의 用語(品詞)가 사이에 있을 때
2) Keyword의 活用語尾가 사이에 있을 때
3) 句讀點, 콜론, 세미콜론, 컴마 등이 사이에 있을 때

또한 連續形으로 하지 않고 漢字 1語로 된 것은 1語를 單位로 하고, 2語 以上인 것은 2語를 單位로 하여 區分하는 方法[28]도 있으며, 複合名詞는 複合抽出하는 例도 있다[29].

(例) 入國拒否 — 入國拒否 / 入國 / 拒否

4.2 Keyword의 選定

收集한 用語中에서 keyword를 選定하는 것으로 다음 事項을 유의하여야 한다.

1) 複數와 單數

한글인 경우는 별 문제가 되지 않으나 英語인 경우에는 다음과 같이 處理한다[30]. 즉, 셀 수 있는 名詞는 複數(gages, nozzles), 物質名詞는 單數(iron, wood, charcoal), 被覆物을 나타내는 것은 複數(plastic coatings), process를 나타내는 것은 單數로(plastic coating), aluminum oxide는 單數, boron orides는 複數로 한다. 그리고 酸化物中 한 種類인 것은 單數로 한다. 其他 詳細한 것은 아래와 같다.

表 4—1. 單數와 複數의 使用法

用 法	單 數 使 用	複 數 使 用
材質 (化合物, 混合物, 物質)	個個의 用語를 表示 : urea, cellophane	包抱的인 用語를 表示 : amines, plastics
性質, 狀態, 特性	viscosity, temperature	physical properties process condition
設備, 裝置, 物體, 素粒子	單數는 使用치 않음	pulverizers, mesons, teeth, stars
用語의 集成	單數는 使用치 않음	adesives, catalysts
프로세스	constructing, modulating	複數는 使用치 않음
固有名詞	hookes, law	複數는 使用치 않음
學科, 分野, 範圍	Chemistry Hydraulics(單數) Engineering	複數는 使用치 않음
現象, 事件	單數는 使用치 않음	ambushes explosions discharges

2) 略 語

略語는 一般的으로 使用하지 않으나, 意味가 確立되어 있는 것은 使用하여도 된다.

(例) UNESCO
KDC
VTOL Aircraft : Vertical take off and landing aircraft
ACTH : Adrenocorticotropic bormone

3) 專門語

專門分野에 따라 차이가 있다. 例를 들어 化合物의 경우는 자주 利用되는 것 以外에는 keyword로 하지 않으며, 특수한 合金은 keyword로 한다[31].

4) 複合語 keyword

하나의 keyword는 그 語數에 關係없이 하나의 槪念을 나타낸다. 그러나 2個 以上의 特定한 keyword를 組合하여 하나의 槪念을 나타내고자 할 때에는 特殊한 複合語로 하든가, 또는 旣存 keyword 2個를 合하든가 하는데, 特殊한 複合語를 만드는 경우의 原則은 다음과 같다[32].

(1) 特殊한 複合語는 適當한 用語가 없을 때에 한하여 만드는 것으로, 槪念을 잘 나타내기 위하여 적어도 하나의 keyword는 그 槪念과 階層的으로 同一한 位置의 一員이어야 한다.

(2) 特殊한 複合語는 그 槪念이 자주 사용되어 檢索이 迅速할 때.

(3) 어느 槪念이 그것을 構成하는 keyword와 階層的으로 同一類였고, 또한 보다 넓은 意味의 keyword와도 同一類였을 때에는 複合語를 使用치 않고 2個 또는 그 以上의 keyword를 各各 使用한다.

以上과 같이 하여 處理된 用語는 字母順으로 排列하여 重複되는 것은 하나로 묶어 頻度數를 記錄한다[33]. 排列이 끝나면 thesaurus에 收錄할 用語를 選定한다. 이 選定作業이 thesaurus 作成의 重要한 作業이며, 다음과 같은 基準을 設定한다.

(1) 用語의 規模를 추정한다. 前述한 바와 같이 專門分野는 약 3,000語이고, 綜合分野는 약 20,000語이다.

여기에서 蓄積文獻數와 索引의 깊이를 추정할 수 있다면 Houston[34]의 式에 의하여 必要한 語數를 豫測할 수 있다.

$$T=3,330\log(P+10,000)-12,600$$

여기서 T : 用語數(keyword)
P : 全文獻數에 대한 全索引語數
$P=Dd$
D : 蓄積할 文獻數
d : 文獻件當 平均索引語數

上記式에 의하여 蓄積文獻 10,000件, 50,000件, 100,000件일 때 文獻件當 索引語 3個씩을 抽出한다면 다음과 같은 keyword數가 나오게 된다.

10,000件일때	keyword는	2,718語
50,000 〃	〃	4,729語
100,000 〃	〃	5,687語

(2) 用語의 使用頻度 데이터가 있으면 그것에 의하여 keyword를 選定한다. 이때 頻度가 높은 것과 낮은 것은 keyword로 選擇하지 않는다. 왜냐하면 頻度數가 높은 것은 어느 文獻에도 빠짐없이 나타나는 것이므로 keyword의 役割을 하지 못하며, 同時에 頻度가 낮은 用語가 主題를 代表할 수는 없는 것이다. 이와 같이 頻度가 아주 높은 用語와 낮은 用語를 keyword로서 取하지 않는 方法은 Luhn[35]의 KWIC 索引法에서 實用化된 것이다. thesaurus의 作成에도 이 아이디어를 適用하는 것이 無妨하리라 생각된다.

4.3 Keyword 關係의 決定

a. 關係語의 決定順序

(1) Keyword를 字母順으로 排列하여 語幹이 同一한 것을 한 곳에 모은다.

(2) 分類表에 따라 keyword를 한 곳에 모은다. 이때 準據할 分類表가 없을 경우에는 별도로 作成하여야 한다. 本硏究附錄에 COSATI의 分類表인 「Subject Cate-

gory Fields and Groups」를 附記하였다.

(3) 위의 方法으로 定하여진 keyword group內에서 keyword 相互關係를 決定한다.

(4) 同形異義語는 ( ) 속에 限定句를 記入한다.

(5) Keyword의 group과 關係없이 各 keyword의 關係를 最終的으로 探知한다.

b. 個個의 關係語를 決定

1) USE와 UF

non-descriptor를 descriptor로 參照하는 것으로 아래와 같은 경우에 使用한다.

(1) 同義語中에서 descriptor로 選定된 것으로 指示

(例) Secondary batteries

USE storage batteries

(2) descriptor로 選定된 것보다 더 一般的인 用語로 參照

(例) Sand blasting

USE Abrasive blasting

(3) 語尾變化된 것중 採擇된 것과 略語에 대한 說明으로부터의 參照

(4) 하나의 槪念을 2個 以上의 descriptor로 나타낼 때

(例) Ferromagnetic film

USE Ferromagnetic materials and film

(5) 同義語로 생각되는 것중에서 descriptor로 採擇된 것으로 參照

(例) Heredity

USE Genetics

(6) 槪念은 同一하나, 보는 觀點에 따라 다른 것중에서 descriptor로 採擇된 것으로 參照

(例) Fluidity

USE Viscosity

(7) descriptor로 採擇된 語順으로 參照

(例) Table(mathematics)

USE Mathematical tables

(8) 現代語를 descriptor로 採用할 경우

(例) Electrical condensers

USE Capacitors

(9) 特殊語를 descriptor로 피할 때

(例) Whirl bird

USE Helicopters

그리고 UF는 non-descriptor를 descriptor로 代表하여 採擇하였다는 것을 參照하는 것으로 반드시 USE에 對應시켜야 한다.

(例) Storage batteries

UF Secondary batteries

2) BT와 NT

階層關係에 있는 用語 하나에서 階層關係 全體를 나타내고자 할 때 使用한다.

(例) Steels

BT Gears

BT에 있어 보다 넓고 보다 좁은 關係는 部分과 全體의 關係와는 一般的으로 다르다. 즉,

Gear teeth

BT Gears

는 옳치 못한 것이다. 그러나 特別한 경우에는 部分 全體의 關係가 成立될 때도 있고, 보다 넓고 보다 좁은 關係가 될 수 없는 경우도 있다. 例를 들면,

Platinum

BT Catalysis

는 되지 못하며,

Platinum

BT Metals

는 될 수 있는 것이다.

그리고 NT는 BT의 逆으로 BT에 모두 對應시킨다.

3) RT

USE↔UF, BT↔NT 以外의 關係를 RT로 定한다. 近似同意語, 서로 部分 全體의 關係, 用途에 의한 關係등 thenaurus를 使用하는 시스템의 觀點에 보아 密接한 關係에 있다고 생각되는 用語는 RT이다.

4.4 編 成

前項까지의 處理가 끝나면 矛盾되는 것이 없나 調査하고 thesaurus format에 맞게 編輯하여야 한다.

a. 矛盾체크

1) 轉置체크

각 keyword는 반드시 對應하는 逆 關係를 이루어야 한다. 이 關係를 調査하는 것을 轉置체크라 하며, 對應關係는 아래와 같다.

A USE B→B UF A

C UF D→D USF C

E NTF→F BT E

G BTH→H NT G

I RTJ→J RT I

2) 矛盾關係체크

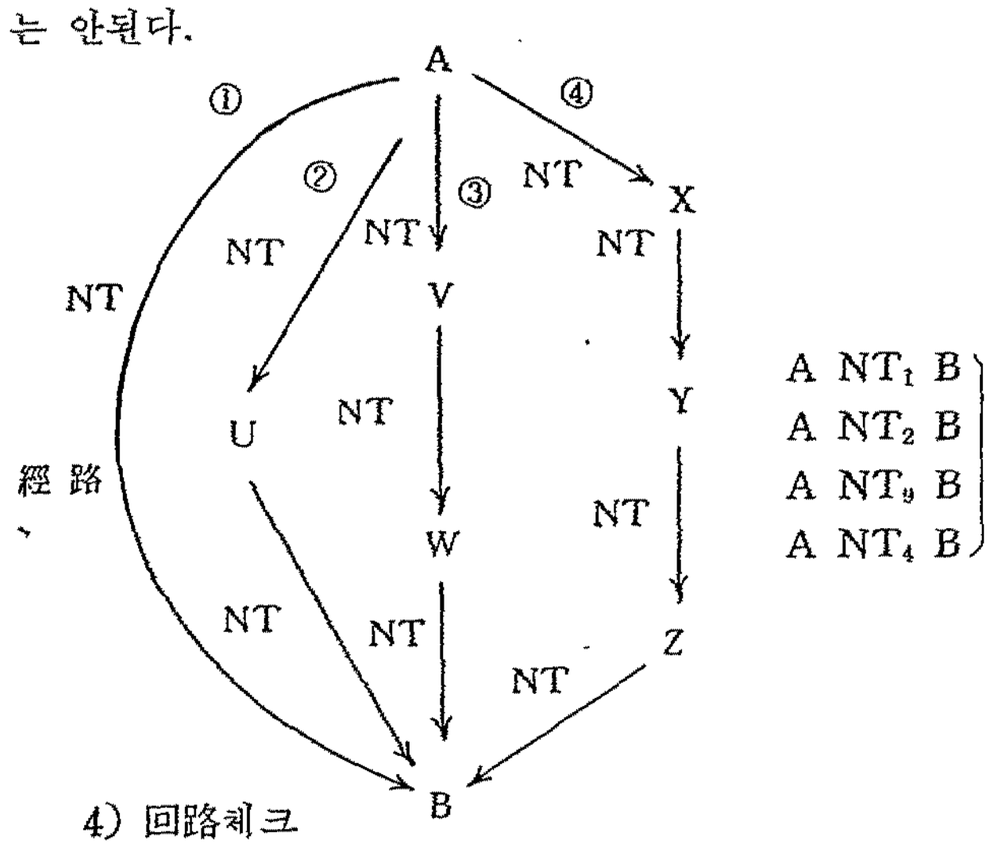
同一한 keyword에 대하여 相異한 關係가 되어서는 않된다.

A NT B
A RT B } 이어서는 않된다.

3) 經路(by pass)체크

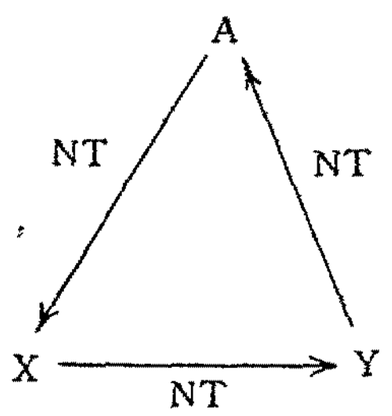
同一 keyword에 대하여 相異한 經路 關係가 있어서

는 안된다.



4) 回路체크

Keyword 自身이 關係를 갖고 있으면 안된다.

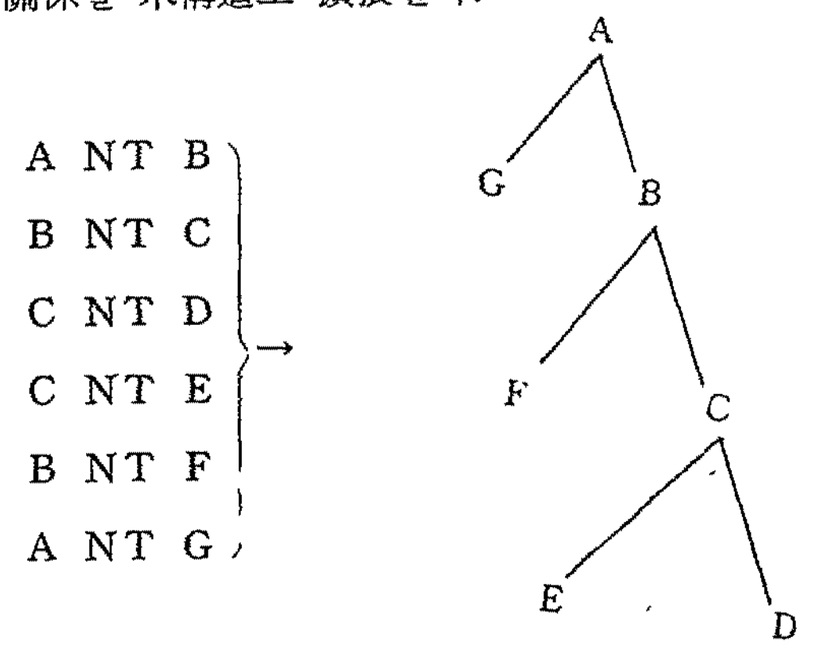


回路關係이다.

A NT_3 A

5) 關係 擴張

階層關係를 木構造로 擴張한다.



b. 編 輯

Thesaurus의 記入을 記入語의 字母順으로 排列하고 記入內를 所定의 規則에 의하여 排列한다. 規則은 關係記號의 順[36], 同一 關係記號에서는 經路順, 여기까지 同一한 것은 關係語의 字母順이 된다. 그 外에 補助的인 事項을 附記하면 記入이 完成된다.

아래의 그림 4—2는 記入의 一例이다.

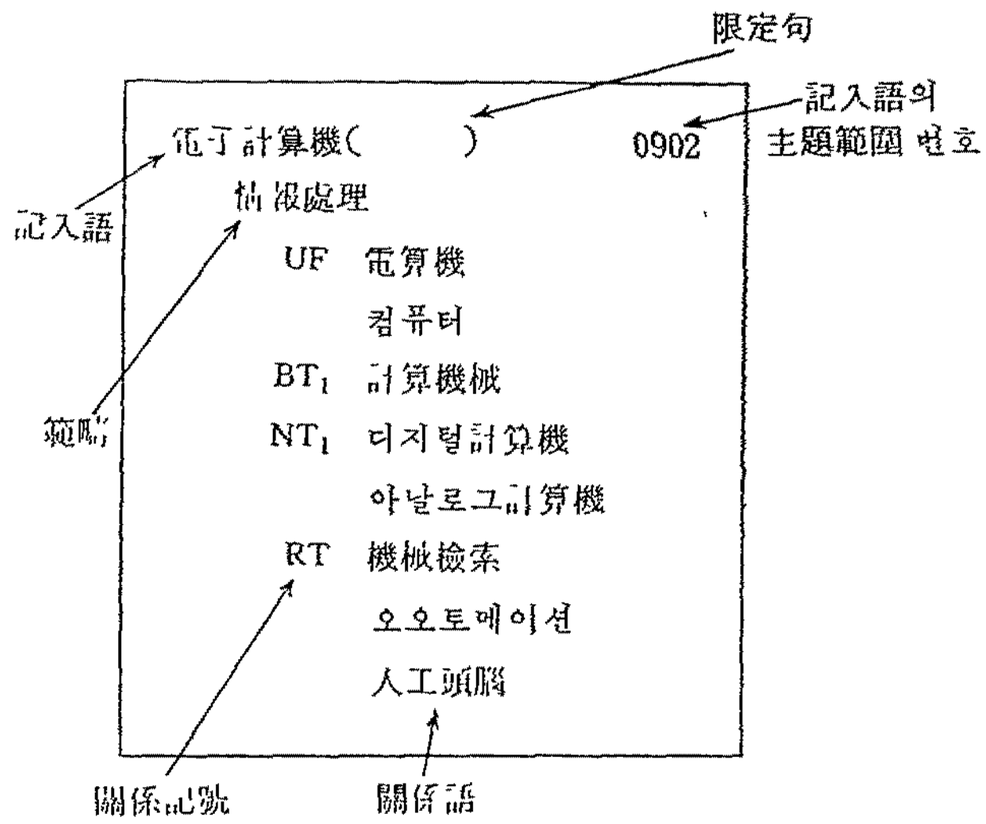


그림 4—2. Thesaurus의 記入例

4.5 更新法

1. 更新의 必要性

Thesaurus는 만들기 위한 것이 아니라 使用하기 위한 것이 아니기 때문에 아래와 같은 理由에서 항상 更新作業이 必要하게 된다.

(1) 學問과 技術이 發達함에 따라 新語가 發生한다.

(2) 同一한 槪念의 體系도 變化하는 일도 있고, 또한 用語間의 意味的 關係도 變化한다.

(3) Thesaurus를 作成한다는 것은 複雜하고 어려운 일이므로 우선 가능한 範圍에서 作成하고 使用하면서 改訂 補完하여 가는 것이 效果的인 方法이다. 그렇기 때문에 Reisner[37]는 使用하면서 繼續的으로 更新(continuous updating)하는 것이 바로 成長하는 thesaurus라 말하고 있다.

2. 更新方法

更新의 對象은 다음과 같다.

(1) 記入語

(2) 關係語

(3) 補助事項

a. 記入語의 更新

一般的으로 收錄되어 있는 모든 keyword를 그대로 남겨 두기를 원하고 있으나, 무조건 남겨두고자 하는 것은 좋은 方法이라 할 수 없다. 使用되지 않는 用語와 앞으로도 使用될 可能性이 없는 것은 削除하는 편이 좋은 것이다. 그리고 新語는 追加한다. 新語라 할지라도 keyword로서 確定치 못할 것도 있고, 또한 自由用語(free term)로 取扱하여 一定期間동안 檢討한 후 採擇與否를 決定하는 方法이 좋다.

특히 自由用語로 取扱할 것은 다음과 같다.

(1) 合金名, 化合物名 그러나 代表的인 合金이나 重要化合物은 keyword로 採擇한다

(2) 個別的 用語 즉, 機種名, 商品名 그러나 thesaurus

의 性質에 따라 keyword로 採擇하는 경우도 있다.

b. 關係語의 更新

한 keyword가 更新되면 당연히 keyword間의 關係도 變化시켜야 한다. 또한 新語가 追加되었다 하드라도 역시 關係는 變化하게 된다. 하나의 keyword에 대한 更新은 連鎖反應的인 影響을 미치기 때문에 手動式에서나 電子計算機 處理에서나 更新은 빠짐없이 實施하여야 한다.

(例)

A NT_1 B NT_2 C NT_3 D	B BT_1 A NT_1 C NT_2 D
C BT_1 B BT_2 A NT_1 D	D BT_1 C BT_2 B BT_3 A

c. 補助事項의 更新

限定句, 記入語의 主題範圍番號 등 記入語의 附隨的인 것을 追加, 改訂, 削除하는 것으로, 一般的으로 記入語를 우선 登錄한 다음에 追加하는 것이다.

參考文獻 및 各註

1) Vickery, B.C. On retrieval system theory. London, Butterworths. 1961. pp. 101~101.

2) 日本ドクメンテーション協会編 シソテス入門(NIPDOKシリース13). 東京, 同協会. 1970. p. 2.

3) Roget's international thesaurus, ed by L.V. Berrey. 3rd ed. New York, Thomas Y. Crowel. 1962로 많이 알려졌음.

4) 日本ドクメンテーション協会編. loc. cit.

5) Brownson, H. Proceedings of the international study on conference on classification for information retrieval. London, Aslib. 1957. pp. 99~100.

6) 化學索引研究グループ. 化學文獻の主題索引法: Thesaurus を用る索引作成實驗. 情報管理. v.8, n.6 1965 p. 3.

7) F.I.D. New Bulletin. Vol. 20, No. 9. 1970. p. 106.

8) Gillum, T.L. Compiling a technical thesaurus. Journal of Chemical Documentation. v.4, n.1 1964. pp. 22~32.

9) Barden, W.A., W. Hammond, and J.H. Heald. Automation of ASTIA; a Preliminary report. ASTIA. AD 22700. 1959.

10) American Institute of chemical Engiheers, Chemical Engineering Thesarus. New York, the Institute. 1961.

11) Science, Government, and Information; a report of the President, science advisory Committee. Washington, D.C., U.S. Government Printing Office. 1963. p. 2.

12) Vickery, B.C. thesaurus; a new word in documentation. Journal of Documentation. Vol. 16, No. 4, 1960. pp. 181~189.

13) 橋本昌幸, 笹森勝之助. ドキユメンテーションにおける辞書の問題 Ⅲ. 情報管理. v. 1, 11, n. 6. 1968 p. 315.

14) Swanson, D.R. Searching natural text by Computer. Science. v. 132, n. 3434. 1960. pp. 1099~1104.

15) Swanson은 "Thesaurus-like Collection of Words and Phrase groups"라 表記하고 있으나, thesaurus라고 해석하여도 무방하리라 생각한다.

16) Lesk, M.E. Performance of automatic information system. Information Storage and Retrieval. v.4, n.2 1968. pp. 201~218.

17) 笹森勝之助. 検索のかなめーシソーラスー. エレクトロクス. v.13, n.11. 1968. p. 1162.

18) Unesco. Guideline for the establishment and development of monolingual scientific and technical thesauri for information retrieval SC/MD/20. 1969.

19) Bernier, C.L. Correlative indexes, Ⅱ Correlative trope indepes, Ⅲ. semanatic relations amaong semantemes. American Documentation. v.8. 1957. pp. 47~50, 211~220.

20) 日本ドクメンラーション協会. op. cit. p. 4에서 再引用

21) Perry, J.W., A. Kent. A tools for machine literature searching. New York. John Wiley. 1956.

22) Committee on Scientific and Technical Information (COSATI). Guidelines for the development of information retrieval thesauri. washington, D.C., U.S. Government Printing Office. 1967.

23) Defense Documentation Center의 略字로 ASTIA가 改稱된 것이다.

24) Schwartz, E.S. Dictionary for minimum redundancy encoding. Journal of the Association for Computing Machinery. v.10, 1963. pp. 413~439.

25) Long, J.M. et al. Dictionary buiding. up and Stability of word frequency in a Specialized medical Orea. American Documentation. v.18, n.1 1967 p. 23.

26) Long, J.M.et al. loc. cit.

27) Engineers Joint Council. Thesaurus of Engineering and Scientific Terms. New York, the Council. 1967.

28) 浜田敏郎. 情報管理におけるシソーラス作成について. びぶるす. v.20, n.10 1969. pp. 1~22.

29) 鈴木幸雄. 外務省におけるシソーラス編成. 第2回 ドキユメンラーション研究集会発表論文集 v.2 1966. pp. 255~258.

30) Engingineers Joint Council. op. cit., pp. 674~675.

31) Engineers Joint Council. op. cit., p. 676.

32) Engineers Joint Council. op. cit., p. 676.

33) 頻度를 調査, 使用하는 것은 收集原이 安定되어 있는 경우이다. 收集源이 多角的이고, 그것도 資料A는 3年分, B는 1年分과 같은 式으로 되어 있을 경우 頻度데이터는 意味가

없게 된다.

34) Houston, N. and E. wall. The distribution of term Usage in manipulative indexes. American Documentation. v.15, n.2. 1964. pp. 111~112.

35) Luhn, H.P. Keyword-in-context index for technical literature. American Documentation. v.11, n. 4, 1960. pp. 285~295.

36) Thesaurus의 種類에 따라 다르나 一般的으로 USE, UF, BT, NT, RT順으로 한다. 이것은 COSATI 方法이기도 하다.

37) Reisner, P. Evaluation of a growing thesaurus. York-town Heights, IBM Watson Research Center. 1967. (Research paper RC—(797)

(p. 140에서 계속)

10. (Canada International Development Research Center(캐나다國際開發研究所)

情報分野의 援助計劃

11. Canada The National Science Library of Canada (캐나다國立科學圖書館)

캐나다식 現況追跡情報서어비스(CAN/SDI)의 Software Package의 提供, 專門職員 派遣, 全國的 SDI 시스템 設置를 위한 訓練提供

12. Mexico 國家 科學技術協議會의 技術情報서어비스(SIT)

產業界 連絡部署의 現場指導

第2回 情報管理研究會 學術大會終了

第2回 情報管理研究會 學術大會가 지난 9월 26일 韓國科學技術情報센터에서 많은 會員이 참석한 가운데 성대히 거행되어 5편의 論文이 發表되었으며, 유익한 질의 응답이 있었다. 그런데 이날 발표된 5편의 論文과 發表者는 다음과 같다.

1. 英國 수개 대학에 있어서의 情報科學 敎育內容에 대하여(金鍾協 : KORSTIC 情報處理部長)
2. 特許審判과 特許出願에 있어서의 特許情報利用(金允培 : 辯理士)
3. 情報處理機械化를 위한 基礎調査—한글, 漢字使用頻度調査—(崔成容 : KORSTIC 調査檢索部 次長)
4. 醫學圖書館에 있어서의 雜誌利用調査(金宗會 : 國防科學研究所 資料管理室長)
5. 대학도서관의 지역사회에 대한 정보봉사—산학협동 체제를 중심으로—(김남석 : 계명대학 도서관장)

原 稿 募 集

情報管理研究會에서는 本紙에 揭載할 原稿를 아래와 같은 요령으로 募集하고 있아오니 여러분의 많은 투고있으시길 바랍니다.

1. 내용 : 정보관리 및 이와 관계되는 내용의 논문, 번역문, 해설기사 등
2. 매수 : 50매내외(200자 원고지)
3. 마감 : 每月 25日
4. 보내실곳 : 131

서울특별시 동대문구 청량리동 206의 9 한국과학기술정보센터내

정보관리연구 편집자앞