Thesaurus의 利用과 最近의 動向

李 正 一
(KORSTIC 기술개발실장)

1. 머리말

Thesaurus란 말이 情報管理·도큐멘테이션의 世界에 登場한 것은 1957年 英國의 Doking에서 開催된 "情報檢索分類에 관한 國際會議"에서 H. Brownson[1]이 행한 講演에서 비롯되었다. 그후 이 말은 여러 專門家들간에 많은 關心事가 되어 왔으며 오늘날 thesaurus 技術의 普及과 發展은 눈부신 바 있다.

thesaurus를 만들면 情報檢索의 問題는 解決되며 thesaurus가 없으면 檢索을 할 수 없다고 速斷하는 傾向이 없지는 않으나 thesaurus가 情報檢索에 있어서 有用한 道具(tool)라는 것을 認識하여, thesaurus 技術을 바르게 利用하고 開發해 나가야 할 것이다.

thesaurus란 희랍어를 語源으로 하여 辭典 百科辭典과 같은 知識의 寶庫를 意味하는 것으로[2] 1852年 P. M. Roget의 "Thesaurus of English Word and Phrases"에서 처음으로 이 用語가 使用되었다. 그러나 情報檢索과의 關係로 話題가 된 것은 前述한 바와 같이 1957年부터이다.

情報管理·도큐멘테이션分野에서 現在, 約 100餘種의 thesaurus가 있는 것으로 推定되고 있지만, 이중에서 가장 일찍 나온 것이 美國國防省의 技術情報센터(ASTIA)에서 作成한 Thesaurus of Descriptors (1960)와 美國化學工學會(AICHE)의 Chemical Engineering Thesaurus (1960)이다. 美國의 工學關係學會의 聯合會인 Engineers Joint Council (EJC)은 1964年에 Thesaurus of Engineering Terms를 發行하였으나 여기에는 Chemical Engineering Thesaurus의 內容이 吸收되어 있다. 1965年 國防省과 工學會聯合會는 各各 改訂할 必要를 느끼고 Project Lex란 共同作業을 開始하여 1967年 Thesaurus of Engineering and Scientific Terms (TEST)이 세상에 나오게 되었다.

Thesaurus에 관한 規格에 대해서는 thesaurus 自體가 情報檢索用語의 標準化의 道具이지만, thesaurus의 作成開發을 위한 國際的 指針이 UNESCO와 ISO TC/46의 共同作業으로 ISO 2788 – 1974 (E)[3]로서 規格化되어 있다. 이 規格의 內容은 분명히 欧美語를 對象으로 하여 만들어져 있으므로 우리말의 thesaurus를 위해서는 우리말의 特性을 充分히 고려한 KS가 必要하다.

또한 이 規格이 成立하게 된 경위나 內容解說에 대해서는 小林氏[4]와 荒木氏[5]의 記事가 있다. 그리고 實用化되고 있는 各種 thesaurus의 作成法, 形式, 機能 및 索引作成 등에 관한 一般的인 解說에 대해서는 司空 哲氏의 "情報檢索論"[6]이나 論文[7]을 參照하면 도움이 될 것이다.

2. 시스템의 類型

thesaurus의 利用이 情報檢索시스템의 效率向上을 위한 重要한 手段이 된다는 事實은 널리 認識되어 있다. 그리고 그 利用方向에 있어서도 2個의 큰 흐름이 있으며, 앞으로 統制語데이터베이스檢索시스템과 自然語데이터베이스檢索시스템의 發展方向에 큰 영향을 미칠 것으로 생각된다. 구체적으로 그 方向을 살펴보면,

(1) MEDLARS 파일用의 MeSH, INIS, ERDA, NASA, JICST thesaurus처럼 統制語데이터베이스檢索시스템에의 利用을 目的으로 한 第1型의 시스템,

(2) INSPEC처럼, 統制語를 主體로 하면서도 디스크립터(descriptor)의 補助로서 非統制語(free term)의 使用을 인정하는 第2型의 시스템,

(3) DDC(美國國防省文献센터)시스템[8]이나 美國 CIA(中央情報局)시스템의 데이터베이스처럼, 統制語는 一部分이고 대부분이 非統制語로 構成된 第3型의 시스템,

(4) IDC(Internationale Dokumentationsgesellschaft für Chemie mbH, Fachinformationszentrum Chemie)시스템[9]처럼, 自然語만 索引하고 "概念그룹化한 thesaurus"를 作成·使用하는 第4型의 시스템의 4가지 方向이 있다.

또한 第1型에도 TEST처럼 科學技術의 全分野를 網羅하는 "Master Thesaurus"的 性格의 것과 教育關係用語를 主體로 한 ERIC thesaurus, 新聞記事用語의 News thesaurus와 같이 使用領域을 明確히 한 것이 있다.

그리고 第4型의 自然語만의 索引에 있어서도 I-DC시스템과 같이 途中에 thesaurus의 同一概念그룹코드로 変換하고 探索에는 自然語의 索引語와 thesaurus의 概念그룹코드의 어느쪽을 選択할 수 있는 시스템, 完全히 自然語로만 索引·検索하는 시스템(피츠버그시스템)[10], 機械援助索引(MAI)과 機械援助探索(MAS)을 目的으로 한 DDC의 NLDA 시스템[11] 등이 있으며 個個의 시스템에 따라 thesaurus도 여러가지로 特徵있는 것이 作成되고 있다.

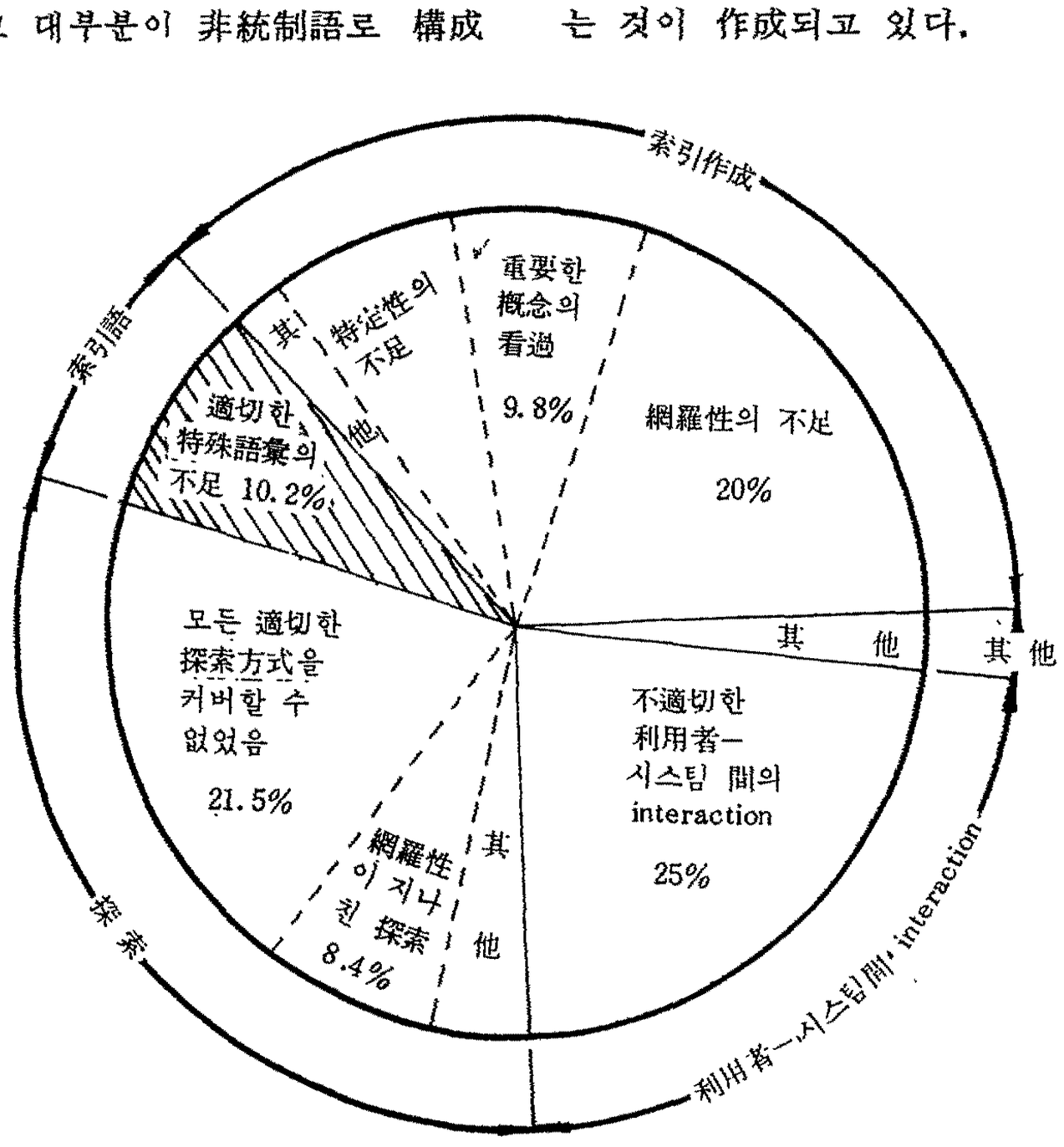

그림 1. MEDLARS에서의 檢索失敗의 主要原因: 再現性失敗의 原因

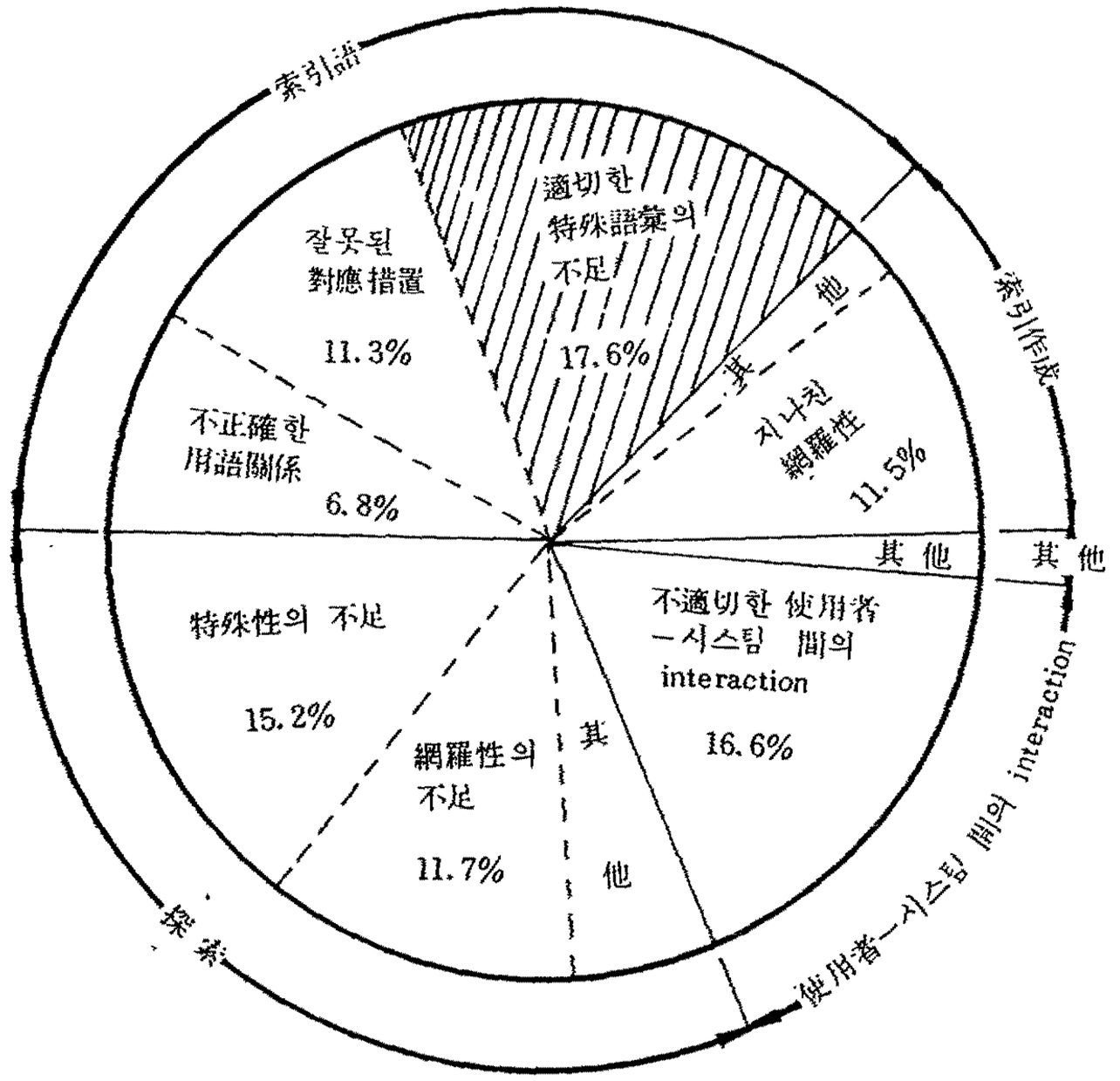

그림 2. MEDLARS에서의 檢索失敗의 主要原因：適合性失敗의 原因

前記의 4가지 型의 시스템에 있어서 共通點은 形式의 차이에도 불구하고, 情報檢索시스템의 效率向上을 위하여 索引作成(入力段階) 또는 探索(出力段階)에서 語彙統制(또는 槪念그룹化)에 의한 thesaurus를 必要로 한다는 것이다. 그리고 4型의 情報檢索시스템은 各各 데이터베이스의 規模의 大小, 包含하는 專門領域 檢索效率을 向上시키기 위한 補助的 手段 등에 있어서 特徵이 있어 일률적으로 比較評價할 수는 없다.

3. 統制語데이터베이스檢索시스템의 特徵과 統制語 Thesaurus의 利用

F. W. Lancaster가 MEDLARS를 對象데이터베이스로 選定하고 評價, 實驗한 結果는 단지 MEDLARS만의 問題가 아니라 統制語 thesaurus를 利用한 統制語 데이터베이스 檢索시스템 全体의 問題로서 重要한 意味를 가지고 있다.

그림 1과 그림 2는 再現性失敗와 適合性失敗의 主要原因을 Lancaster가 分析, 提示한 것이다.

그림 1, 2에서 알 수 있는 바와 같이 索引語 및 索引作成上의 잘못이 再現性 失敗, 適合性 失敗의 48%를 차지하고 있다.
즉 索引語 및 索引作成에 起因하는 失敗가 結局, 시스템全體에 대한 效率에 크게 影響을 미치고 있다. 상세히 檢討하여 보면 索引의 特定性(Specificity)의 不足에 의한 失敗도 索引語의 特定性의 不足과 階層關係·關聯語의 表示上의 問題 등 統制語 thesaurus가 關與하고 있는 경우가 대단히 많다.

統制語 thesaurus 중에 없는 새로운 槪念의 語彙, 더구나 特定性이 있는 語彙는 앞으로 계속 增加할 것이므로 索引의 一貫性을 維持하고 探索에 有用하게 再現性을 높이고자 하면 할수록 統制語는 그만큼 增加될 것이다 뿐만 아니라 시스템 維持經費를 增大시키고 索引作業의 迅速性이 低下하는 結果를 초래하게 된다.
여기에 統制語 thesaurus의 限界가 있다고 생

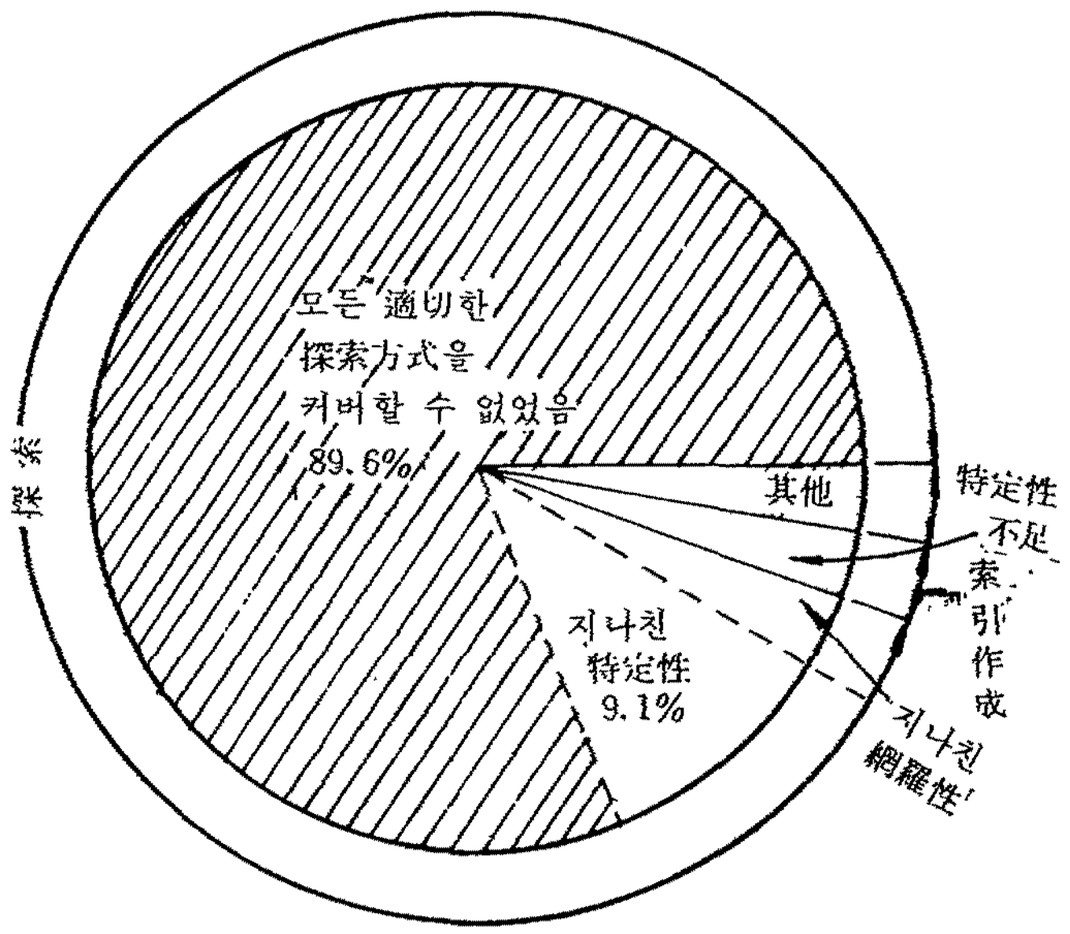

그림 3. EARS(自然語데이터베이스)에서의 檢索失敗의 原因：再現性 失敗의 原因

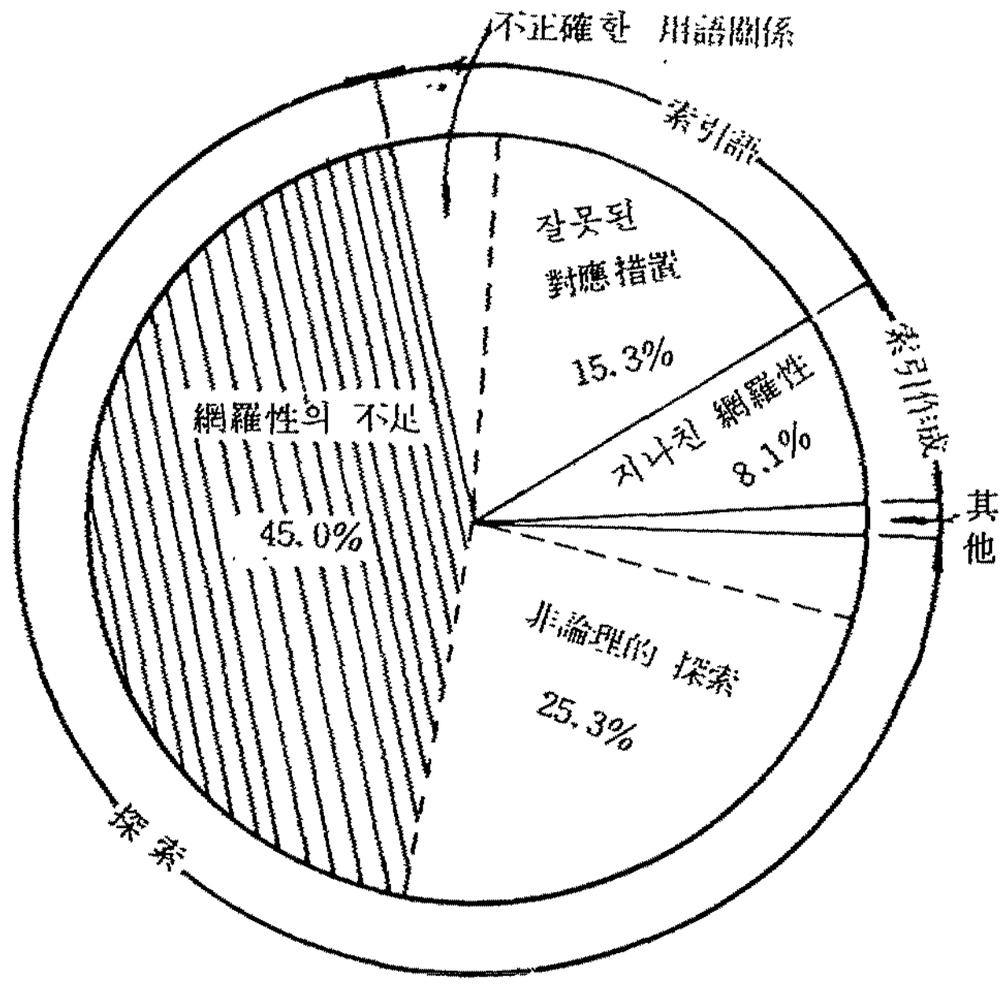

그림 4. EARS(自然語 데이터베이스檢索시스템)에서의 檢索失敗의 原因：適合性 失敗의 原因

각된다. 또한 適合性 向上을 위한 補助的 手段으로서 Link와 Role이 使用되고 있다. 그러나 Role의 利用에 관해서는 一部의 特殊한 領域에만 限定되어 있고 大規模시스템에의 應用에 관해서는 經費나 效率上의 問題가 있다는 것이 實驗的으로 指適되고 있다.

統制語 thesaurus의 또하나의 重要한 點은 thesaurus의 "構造와 表示法"에 관한 것이다. 다시 말하면, 索引 또는 探索에서 어떤 槪念에 관한 用語를 생각하면서 그 用語가 어떤 디스크립터로 使用되고 있을까, 더구나 디스크립터가 없는 경우에 어떤 用語를 使用하면 좋을 것인가를 迅速, 正確하게 決定할 수 있는 機能을 가지고 있을까 하는 點이다. 이것에 대해서는 TEST thesaurus와 Roget－Soergel model thesaurus를 使用하고 比較分析한 例가 있다.[12]

이것과는 달리, "統制語＋非統制語 (free term)" 方式으로 索引語의 特定性의 不足을 보완하고 있는 INSPEC, Open－ended term 으로 統制語의 몇배의 非統制語를 計算機파일 중에 索引時에 入力하고 必要에 따라 端末에서 探索에 使用할 수 있도록 되어 있는 DDC시스템 및 IDC thesaurus에 있어서 自然語索引과 槪念코드의 利用 등은 自然語와 統制語의 特徵을 살린 注目할 만한 方向이라 생각된다.

4. 自然語데이터베이스檢索시스템의 特徵과 探索語 Thesaurus의 利用

F. W. Lancaster가 EARS (Epilepsy Abstracts Retrieval System)를 對象데이터베이스하여 評價實驗한 結果에 의하면[13] 自然語데이터베이스의 特徵을 잘 알 수 있다. 適合性과 再現性의 失敗의 原因은 그대로 自然語 데이터베이스의 對策의 方向을 暗示하여 주고 있다.

그림 3과 그림 4는 失敗의 主要原因을 提示한 것이다.

그림 3, 4에서 알 수 있는 것처럼 自然語를 基本으로 한 온라인시스템에서는 檢索失敗의 原因은 데이터베이스(索引語나 索引作成)에 있는 것이 아니라 데이터베이스를 探索하기 위하여 使用한 探索方式에 그 原因이 있다.

自然語 데이터베이스에서는 索引語는 完全한 特定性이 있으며 探索은 自然語와 거의 같은 수준에서 행할 수 있다. 그러므로 再現性 失敗의 95%, 適合性 失敗의 30%는 그의 探索戰略에 原因이 있는 것이다. 따라서 自然語 데이터베이스의 探索者는 探索을 包括的으로 수행

하고, 再現性을 높이기 위해서는 探索을 效果的으로 하는 어떤 補助的 手段이 必要하다. 이를 위한 하나의 手法이 thesaurus의 利用이다. 이 경우의 thesaurus는 自然語의 特徵을 살린 探索用 thesaurus이다.

探索用 thesaurus의 實例로서는 法律關係에 實用化되고 있는 피츠버그시스템의 thesaurus[14]가 있고, SMART(實驗的 自然語 自動索引 데이터베이스 檢索시스템)의 thesaurus가 있다.
또한 보다 간단한 예로서는 T. Whitehall (Brooke Bonde Liebig Research Centre)[15]이 提案하고 있는 "利用者를 위한 thesaurus", "Mini thesaurus", "個人 thesaurus" 등 目的에 따라 여러가지 形態로 探索用 thesaurus가 作成되고 있다.

自然語 데이터베이스에 있어서 하나의 重要한 方向은 包括的인 槪念의 主題코드로 된 小規模 統制語와 그것에 關聯한 自然語索引이란 2段階索引·探索方式이다.
이러한 것의 예로서는 美國 CIA의 시스템, 原子核科學分野의 thesaurus[16] 등이 있다.
自然語의 特定性을 살린 索引을 하고, 主題코드로 어느 程度까지의 索引의 一貫性을 確保하며, 探索에서는 統制語主題코드로 主題領域을 限定하여 自然語探索에서 그의 特定性과 網羅性(exhaustivity)을 높이고자 하는 것은 注目할 만한 方向이라 생각된다. 이 "2段階探索法"의 評價는 분명하지는 않지만, 大規模 統制語 데이터베이스의 thesaurus構造의 高度化, 統制語의 增加 및 索引者의 教育 등에 對應하고자 노력하는 現在의 趨勢와 더불어 情報檢索시스템의 簡略化에의 하나의 움직임으로서 注目해야 할 것이다.

5. Thesaurus의 動向

1960年 5月의 ASTIA thesaurus 第1版으로 始作된 60年代의 "統制語彙" 時代로부터 "Free Term의 追加" 時代에로의 移行내지는 DDC의 NLDA나 SMART로 代表되는 Free Term派의 대두의 조짐이 엿보이고 있다.
이 傾向은 ASTIA thesaurus를 使用하면서도 Open-ended term 및 identifier의 導入으로 索引된 用語中에서 thesaurus의 用語가 全體의 5% 前後로까지 低下한 DDC의 現狀이나 thesaurus語彙에 free term을 追加하여 索引하고 있는 INSPEC('71 ~)의 狀況을 보아도 명백하다. 또한 實際로 利用하고 提供되고 있는 데이터베이스 및 온라인檢索시스템에 free term 追加機能을 가진 것이 數量的으로 伸張되고 있는 것도 하나의 추세라 할 것이다.

完全한 自由語彙派인 BIOSIS에서는 5,000,000語가 넘는 방대한 語彙와 그의 探索에 있어서의 費用 때문에 頻度가 높은 語의 周辺에 대하여 整理하는 方向으로 나아가고 있다.[17] 단 어디까지나 自由語彙派이므로 自然語로 索引, 蓄積한 것을 整理하고 있는 것이다.

完全統制語彙派인 Lancaster도 統制語彙에 의한 特定性의 결함을 보완하기 위하여 entry 語彙를 提唱하고 索引作業의 槪念分析段階에서 重要時해야할 語나 句는 모두 entry語彙에 包含시킬 수 있다고 發言하고 있다. 그러나 그는 어디까지나 完全統制語彙派임을 고수하고 있다.

또하나의 傾向으로서는 複數의 데이터베이스

```
0902
Computers
  Accumulators (computers)
  Airborne computers
  ALGOL
  Analog computers
  Analog to digital converters
  Aperture cards
  Arithmetic and logic units
  Assembler routines
  Assembly languages
  Associative storage
  Asynchronous computers
  Autocoders
  Auxiliary equipment (computers)
  BASIC (programming language)
  Binary processors
  Bombing computers
  Buffer storage
  Calculators
  Card punches (data processing)
  Card readers (data processing)
  Card reproducers
  Card sorters
  Card to tape converters
  Central processing units
  Character generators
  Character processors
  Character recognition devices
  COBOL
  COGO
  Collators
  Compilers
  Computer components
  Computer driven punches
  Computerized simulation
  Computer logic
```

그림 5. 카테고리索引(TEST)

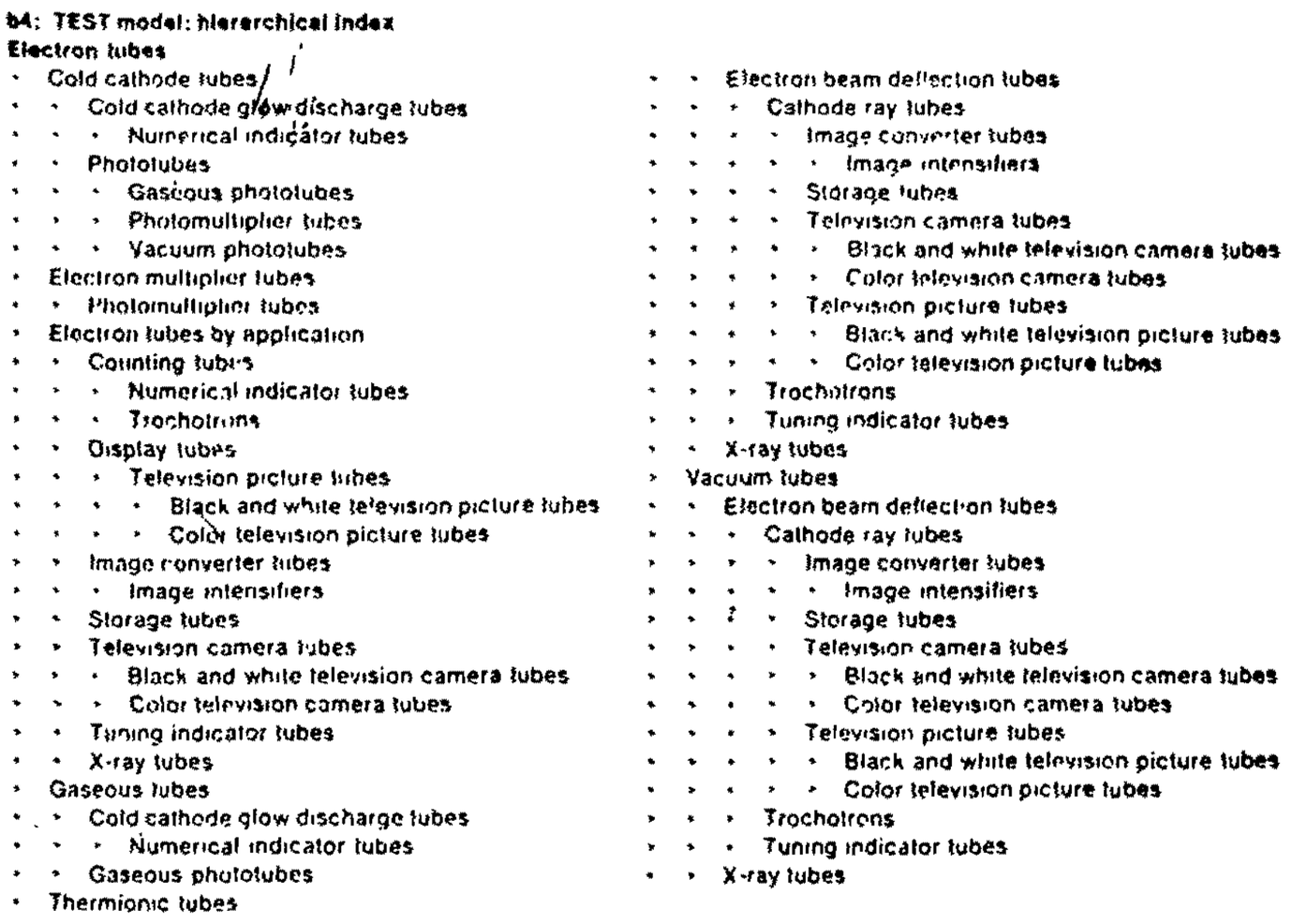

그림 6. 体系索引(TEST)

(探索用 파일)가 쉽게 利用될 수 있는 多重레이 터베이스時代의 必然的인 要請으로서 利用者는 여러 種類의 語彙(thesaurus도 包含)를 알지 않으면 안되게 되었다. 그러나 豊富한 經驗을 쌓은 NTIS(美國國立技術情報서비스機關)의 專門家조차도 4種類의 thesaurus를 驅使하지 않으면 探索이 어렵다고 Urbach('73)는 報告하고 있다.

一般利用者에게 各種의 thesaurus를 熟知토록 要求하는 것은 무리이다. 그 結果 解決策으로서 "thesaurus의 整合性"이 要請되고 있다.

또한 情報의 生産者, 加工者, 利用者를 分業하는 것이 아니라 同一人物이 實行한다는 情報시스템의 個人所有指向을 들 수 있다. 汎用化, 大型化라는 것에 결코 異端者로서가 아니라 그것과 연결하면서도 個人 thesaurus, 個人시스템으로서의 研究者파일에는 情報処理시스템을 신변 가까이 完全한 것으로 하기 위한 未來가 있을 것이다.

끝으로 시스템의 語彙構造 그것만이 아니라 thesaurus 利用面에서 본 構造形態, 즉 語彙目録에의 探索容易表示에 관한 것이다. 統制語彙나 自由語彙나하는 論議는 探索主題에 適合한 用語의 集合을 즉시 그리고 간단히 列擧할 수 있느냐의 問題로 바꾸어 말할 수 있다.

語彙目録에서 必要한 用語를 모두 찾아낸다는 것이 긴요하다. 現在의 語彙目録은 알파벳順리스트와 thesaurus(階層)리스트로 大別할 수 있다. 前者에 대해서는 그 用語가 속하는 카테고리나 主題別로 나눈 카테고리索引(그림 5)과 主題索引, 體系索引(그림 - 6)이 있지만 어느 것이나 前方一致機能만 가지고 있다. 그러므로 中間一致, 後方一致의 機能을 가진 것이 바람직하다. 後者(階層)리스트에 대해서는 對象分野에 따라 弾力的으로 고려해야 할 것으로 생각된다.

6. 맺는말

지금까지 情報検索시스템全體中에서의 thesaurus의 利用이란 觀點에서 實驗例를 基礎로 하여 thesaurus의 機能을 紹介하고 그 動向을 개략적으로 살펴 보았다.

關心있는 분에게 다소나마 도움이 되었으면 한다.

參 考 文 獻

1) Brownson, H. Proceedings of the International Study on Conference on clasification for Information Retrieval.
London, Aslib, 1957. pp. 99～100.

2) 日本 ドキュメンテーション協会編. シソーラス入門' 東京, 同協会, 1970. p. 2.

3) ISO 2788－1974 (E), "Documentation－Guidelines for the establishment and development of monolingual thesauri", pp. 13.

4) 小林和雄, "シソーラス作成についての ユネスコ案・ドイツ案の 比較", 情報管理. Vol. 15, №1, 1972, pp 24～32.

5) 荒木啓介. "単言語シソーラスの 設定と發展のための 指針－ISO 2788 1974年版について", 情報管理. Vol. 19, №5, 1976, pp 343～349

6) 司空 哲. 情報検索論. 서울, 亜細亜文化社, 1977.

7) 司空 哲. "情報検索에 있어서 Thesaurus 의 導入에 관하여", 情報管理研究, Vol 7, №5, 1974 pp 115～125.

8) Klingbiel, P. H., "The future of indexing and retrieval vocabularies", 1970. DOC－TR－70－4, AD 716200.

9) 長谷川正好, etal, "IDC シソーラス", 情報管理, Vol 20, №10, 1978, pp. 807～817.

10) Lauren D. B., "Semantic Road Maps for Literature Searchers", J. ACM Vol 8, №4, 1961. pp. 553～578.

11) Lancaster, F. W., MEDLARS; Report on the evaluation of its operating efficiency, American Documentation, 1969. pp 119～142.

12) Soergel, D.: Indexing Languages and Thesauri;

Construction and Maintenance, Melville Puhlishing Company, Los Angeles, 1974

13) Lancaster, F. W. etal., "Evaluating the effectiveness of an on－line, natural language retrieval system", Information Storage and Retrieval, 8, 1972. pp. 223～245.

14) Lancaster, F. W.: Vocabulary Control for Information Retrieval. Information Resources Press. Washington, 1972.

15) Whitehall, T, "A thesaurus for the user", Conf.
Informatics 1, 1974. pp. 135～144.

16) Keen E. M., "The Aberystwyth Index Languages,"
J. Doc. Vol 29, №1, 1973. pp. 1～35.

17) Lefever, M. etal., "Managing an Uncontrolled Vocabulary",
J. Am. Soc. Inf. Sci, Vol 23, №6, 1972. pp. 339～342