農学系 索引語彙에 관한 硏究

—自動化를 위하여—

이 태 영
(숭의여자전문대학 강사)

Ⅰ. 서 론

1. 硏究의 目的과 意義

컴퓨터가 情報資料의 蓄積, 統制, 檢索에 利用된 以後, 資料檢索 분야에서는 전통적인 수작업시스팀에서 볼 수 없었던 양과 質을 겸비한 광범위하고 신속하며 정밀한 探索을 실시할 수 있었으나 우리나라에서는 아직까지 전산화 情報檢索시스팀이 보편화되고 있지 못한 실정이다.

특히 農學관련 분야는 産業經濟技術硏究院에서 AGRIS Coordinating Center[1]와 바터제로 위촉 운영되는 AGRIS 데이터베이스파일만이 그 名目을 유지하고 있을 뿐 韓國의 農學유관기관에서 自體로 製作하여 사용하는 檢索시스팀은 없다.

그런데 農學분야는 先進技術國들의 첨단을 걷는 學術硏究와 더불어 지역적인 硏究가 한층 중요시되는 硏究領域이므로 신속하고 정확한 國內外 學術資料의 情報流通시스팀이 필요하다. 따라서 韓國語, 英語, 日語資料등 農學硏究者들의 硏究에 없어서는 안될 情報資料를 쉽게 探索하여 情報원을 획득할 수 있게 해주는 혼합데이터베이스가 요청된다. 英語와 日語資料의 데이터베이스는 그 해당국과의 협약에 의해 이용할 수 있으나 韓國語資料는 우리自體內에서 손수 製作하여 이용하지 않으면 안된다.

또한 韓國語, 英語, 日語에 관한 同義語사전을 소장시켜 어느 한나라말로 探索할 때에도 해당되는 3개국어자료가 동시출력될 수 있도록 시스팀을 개발하면 시간과 경비절감에 많은 도움을 줄 수 있다. 이러한 檢索시스팀을 개발하고 그 운영을 成功的으로 수행하기 위해서는 檢索시스팀의 核心인 索引言語가 우선적으로 硏究되어야 한다.

특히 電算化 情報檢索시스팀의 長點인 후조합색인시스팀의 키워드색인어휘의 開發과 어휘집 및 사전의 편찬이 중요하므로, 農學系索引어휘들의 特徵과 속성을 分析하고 이들의 索引言語를 발췌하는 수법을 개량하여 효용성있는 語彙的 모형을 提示하는데 본 연구의 目的이 있다.

2. 硏究의 범위와 方法

혼합데이터베이스 운영에 관한 본 索引言語 硏究는 英語와 日語에 관한 問題에 있어서는 각국의 硏究 結果를 이용하면 되므로 우리가 해결해야할 韓國語 索引語彙에 국한시켰으며 韓國語 기사에 出現하는 英語용어들에 관한 問題만을 다루었다.

硏究方法은 시스팀設計의 目的으로 農學中 여러 分野에 접촉이 된다고 인정되는 植物병리학

1) 국제적인 농학정보 유통기관

을 택하고, 利用者층은 대학생 이상 수준으로 설정하여 시스템分析을 하였다. 즉 分析統計値 및 索引語 抽出을 위해 韓國作物學會誌, 韓國植物보호학회지, 韓國원예학회지, 韓國균학회지, 韓國축산학회지, 韓國農化學會誌, 農村經濟, 韓國農工學會誌의 記事들과 農學系디소러스인 "AGROVOC"[2], "Agricultural Term"[3], "Agri-Cultural/Biological Vocabulary"[4], "農業英語集"[5], "한글농업용어집"[6], "식물병리학"[7]을 참조하였고, 특히 索引言語의 효과적인 分析을 위해 言語學에서 다루는 文法의 기초를 이용하여 農學用語들의 실상을 糾明한 후에 索引語彙集(디소러스)작성에 참고하였다.

索引語 발췌와 어휘집 형성에 편리를 기하기 위하여 실험적 이론인 自動索引발췌와 어휘집 형성론 중 Harter와 Salton의 방법을 응용 실험하였다.

II. 理論的 배경

1. 索引用語의 고찰

索引은 다변적인 개념을 내포하고 있다. 처음에는 단순한 単語索引의 기능을 수행하였으나 19C이후 主題索引으로 그 역할이 확대되면서[8] 부터 索引은 목록의 기능을 갖게 되었으며 표목으로 올려지는 표목어들의 형태에 따라 分類, 主題名, 저자명, 키워드색인 등으로 지칭되며, 이들 표목으로 올려지는 것을 통괄하여 索引言語로 규정한다.[9]

索引言語의 하위개념으로 색인어휘, 색인어, 의미어,[10] 코드, 보조소, 형성규칙이 있다.[11,12] 索引語彙는 索引語의 집합적인 뜻으로 사용이 되며, 索引語는 색인표목으로 채택되는 개개 単語를 말하는 것으로 선택어(descriptor)로도 불리워진다. 의미어는 索引語로 선택되어질 수 있는 후보용어들을 총체적으로 지칭하며, 보통 문법적으로 명사, 명사형 중 주제를 강하게 띄고 있는 단어들이 포함되고, 코드는 종래의 기호를 의미하는데, 이는 어휘를 특정 기호로 표현해 놓은 변환표시를 말한다. 그리고 보조소와 형성규칙을 보면, 전자는 색인어를한정하고 수식하는 역할을 하는 역할기호, 연결기호, 한정어 등을, 후자는 색인어의 형태를 구체적으로 지시, 색인어간의 관계를 명시하는 일련의 항목과 그 부수기호들을 일컫는다.

2. 索引의 理論

색인이론은 이차대전후 폭발적인 정보량의 증가로 말미암아 정보처리문제로 각국이 고심할때 검색효율을 높이기 위해 차원 높은 기

---

2) AGRIS Coording Center, AGROVOC, Rome; The Center, 1981.
3) National Agricultural Library, Agricultural Terms, Phoenix; Oryx Press, 1978.
4) National Agricultural Library, Agricultural/Biological Vocabulary, Washington, The Library, 1967.
5) 농촌진흥청, 農業英語集, 수원; 농촌진흥청, 1978.
6) 농촌진흥청, 한글농업용어집, 수원; 농촌진흥청, 1971.
7) 정후섭, 植物病理学, 서울; 향문사, 1980
8) Encyclopedia of Library and Information Science, Vol. 11(1971), pp.268-299.
9) J. L. Jolley, "The terminology of coordinate indexing", Aslib Proceedings, Vol. 28, No. 3 (1976), pp. 120-128.
10) S. P. Harter "A probabilistic Approach to Automatic Keyword Indexing: part 1 on the Distribution of specialty words in technical Literature", JASIS, Vol. 26, No. 4(1975), pp. 197-206.
11) F. W. Lancaster, Vocabulary Control for Information Retrieval, Washington; IRP, 1972, pp. 115.
12) D. Soergel, Indexing Languages and Thesauri: Construction and Maintenance, LosAngeles; Melville Pub. Co., 1974. pp. 27-28.

술과 이론을 개발하기 시작한데서 비롯된다.[13]

1950년대 Jonker와 Heilprin은 각각 2원적 모형, 3원적 모형을 제시하였으며,[14] 1970년대까지 확률이론, Landry와[15] Salton의 이론, 효용이론, 확률효용이론[16] 등이 대두하였는바, 본고에 응용한 이론은 Landry, Salton, 확률이론이다.

2.1 Landry의 理論

Landry의 이론은 색인작성과정과 일반통신과정을 유사하게 보는데서 출발한다.

한 단위의 통신과정이 각각의 통신부호의 흐름이듯이 한 문헌의 색인과정도 문헌내 단어들의 흐름중에서 일정 단어군을 취택하여 잘 정열시킨 것을 색인으로 간주한다.

그는 자료소의 개념을 ①독립적존속가능, ②분해할 수 없는 최소단위, ③정의되어질 수 있는 의미를 갖는 것으로 정의 하였다.

2.2 확률이론[17),18),19)]

문헌 내에 분포하는 단어들의 출현현상을 일정한 확률통계법칙에 적용하고 색인어와의 상관관계를 유추하는 것이 확률이론이다.

1960년대 전반 C. T. Abraham 등이 제시한 Poisson분포이론은 초기확률이론의 대표적인 것이었다. 통계적으로 일정 문헌군의 단어들의 출현 현상을 검토한 후 그 현상이 Poisson분포를 닮았다고 설명하였다. 즉 Poisson 분포는 일정 문헌군내의 각 단어의 출현확률 $\lambda = \frac{wi}{A}$ (wi→각 단어 출현빈도수, A→문헌수)가 정해지면 Poisson함수공식에 따라 각 단어 wi가 몇개 문헌(di)에서 몇번씩(ki) 출현할 것인가를 예언하는 것이다. 그후 Bookstein, Harter[20] 등은 Poisson분포는 무의미어(조사, 관사와 같은 기능어)들에 적용이 되고 의미어는 2-Poisson분포를 따른다고 주장하였다.

Harter의 실험결과, 실제로 의미어는 단일 Poisson (Single-Poisson)에 따르지 않았으며 2-Poisson에 어느 정도 합치되었다. 2-Poisson은 $P(k) = \pi \frac{e^{-\lambda 1}\lambda_1^k}{k!} + (1-\pi)\frac{e^{-\lambda_2}\lambda_2^k}{k!}$ (여기서 π와 λ_1, λ_2는 moment generation function으로 구할 수 있다.)와 같이 Poisson 공식을 수정한 것인데, Harter는 이 공식에서 의미어와 무의미어를 구별하기 위하여 Z치를 고안하였다. 이 Z치는 그 값이 낮으면 무의미어, 높으면 의미어로 구별을 시켜주는 식별치이다.

2.3 Salton의 이론

Salton은 자동어휘집 형성에 관하여 몇가지 방법을 이론 및 실험적으로 제시하였다.[21,22] 그의 제1방법은 완전자동기법으로서 용어집합

13) Encyclopedia of Library and Information Science, op. Cit.

14) H. Borko, "Toward a Theory of Indexing", Information Proceedings & Management, Vol. 13, (1977), pp 355-365.

15) B. C. Landry, "Toward a Theory of Indexing-II", JASIS, Vol. 21, NO. 5(1970), PP 358-367.

16) W. S. Cooper, "Foundations of Probabilistic and Utility-Theoretic Indexing", JACM , Vol.-25, No 1, 1978. pp. 67-80.

17) A. Bookstein, "Probabilistic Models for Automatic Indexing, "JASIS, Vol. 25, No5, (1974) pp. 312-318

18) Maron, M. E., "On Relevance, Probabilistic Indexing and Information Retrieval, "JACM. Vol. 7(1960) pp. 216-244.

19) Robertson, S. E., "The Probabilistic Character of Relevance, ", Information Processing & Management. Vol. 13(1971) pp. 147-151.

20) S. P. Harter, op. cit.

21) G. Salton, Automatic Information Organization and Retrieval, New York; Mc Graw-Hill Co, 1968. pp49-64.

22) G. Salton, "Experiments in Automatic thesaurus Construction for Information Retrieval", Sci. Rep. Inform. Stor. Retr. NSF. Vol. 18, (1970). pp. Ⅶ-1~Ⅶ. 28.

V의 각 단어들, wi가 일정문헌군집합, D의 각 문헌, di에 출현하는 빈도수를 V와 D간의 행렬값으로

	w_1	w_2
d_1	2	1
d_2	1	2

(D와 V의 요소가 각각 2개일 경우)와 같이 배정한 후 각 단어 wi를 $w_1 : w_2$, $w_2 : w_1$, 의 순으로 비교하는데 단어 wi의 벡터값(행렬값) d_1, d_2,를 같은 순위끼리 매칭시켜 최소치를 선택하여($w_1 : w_2 \rightarrow d_1=1, d_2=1$)합산한다. 이때에 w_1을 중심으로 w_2를 비교하였을 때에는 w_1의 벡터값을 합산하여

$\frac{\Sigma\, min(w_{12})}{\Sigma\, w_1}$ 를, w_2를 중심으로 비교하였을 때에는 $\frac{\Sigma \min(w_{12})}{\Sigma\, w_2}$ 를 단어 대 단어행렬

	w_1	w_2
w_1	1	2/3
w_2	2/3	1

에 행렬값으로 배정한다. 다음에 임의의 식별치 γ을 결정하고

1) $w_{12} > \gamma$, $w_{21} > \gamma$일 때는 동의어 관계
2) $w_{12} > \gamma$, $w_{21} < \gamma$일 때는 w_1이 w_2의 상위어
3) $w_{21} > \gamma$, $w_{12} < \gamma$일 때는 w_2가 w_1의 상위어
4) $w_{12} < \gamma$, $w_{21} < \gamma$일 때는 관계가 없다고 결정한다.

두번째 방법은 반자동식으로서 일련의 性質項(Properties) 집합P를 설정하여 각 성질항 Pi를 선정하고 선택된 각 단어들 di가 성질항에 어느 정도 적응도를 갖는가를 판별하여 일정값을 부여한다. 이들의 관계결정 절차는 P를 D로 대체한 상태의 첫째방법과 똑같은 수순으로 정리한다.

세째방법은 군집화(Clustering)를 컴퓨터의 Sorting능력을 빌어 수작업으로 처리하는 방법이다.

3. 농학계 기존색인 시스팀고찰

3.1 電算化 情報検索시스팀

전세계적으로 이용되고 있는 농학분야의 데이터베이스는 AGRIS[23], AGRICOLA[24], CAB[25], BIOSIS[26]가 있다. 이 데이터베이스들의 내용 및 特徵을 잘 나타내는 項目을 표 1에 要約

표 1.

	AGRIS	AGRICOLA	CAB	BIOSIS
供給機関	FAO	미국 농무성	영국 농림국	미국 생물공학정보서비스
索引言語	통제언어 : 영어	통제, 자연언어 : 영어	통제, 반통제, 자연언어 : 영어	비통제 : 영어
分類코드	주제범주코드 개체코드	주제범주코드	광범위주제범주코드	주제개념코드 생물분류코드
検索코드	색인어, 저자명 주제범주코드, 표제, 언어코드, 지명	기본색인, 일반색인, 저자명, 표제, 언어코드, 주제범주코드	기본색인, 일반색인, 기관명, 저자명, 잡지명, 지명, 언어코드, 표제, 주제범주코드	색인어, 저자명, 표제, 주제개념코드, 생물분류코드
색인어 특정성	아주 포괄적 또는 극도의 세분어휘 제외	=	=	=

23) F. W. Lancaster, "Assessing the Benefits and Promise of an International Information Program(AGRIS)", JASIS, Vol. 29, No.6(1978), pp. 283－288.
24) Charles L. Gilreath, "AGRICOLA; Multipurpose Data for Agricultural and Life Science Libraries", Serials Librarian, Vol. 3, No.1(1978), pp. 89-9 5
25) Blaise Cronin, "CAB Abstracts; A global View", Aslib Proceedings, Vol. 32(1980). pp. 425－437.
26) 森岡祐二, " BIOSIS(生物)", 情報管理, Vol. 23, No.8 (1980), pp. 721－736.

紹介한다. 전산화시스팀의 특징은 多様한 検索코드에 있으며, 특히 基本索引과 一般索引部를 갖는 後組合索引에 따른 探索方法이다.

3.2 国外 索引法

국외 발행 索引誌로서 국내에서 이용되고 있는 冊子型 索引誌는 前記한 AGRIS, AGRICOLA CAB, BIOSIS에서 機関誌로 발행한 "AGRINDEX", "Bibliography of Agriculture", "CAB Abstracts", "BA/RRM"이 대표적이며, 그외에 "Biological & Agricultural Index"와 "AGRONOMY" 등이 있다. 이들 国外 索引誌는 電算化시스팀내에 収録된 레코드의 項目과 거의 同一하며, 冊子構成에서 "주기입부문", "著者索引", "主題索引(分類索引)", "지리색인"별로 편집이 되어 検索코드는 著者名, 分類코드, 地理코드, 標題를 활용할 수 있다.

3.3 国内 索引誌

과거에 出刊되었거나, 현재 出刊中인 것을 網羅하여 農業分野文献을 포함하고 있는 국내 색인지로는 "国内定期刊行物記事索引", "農村振興事業文献情報", "국내과학기술연구과제총람", "韓国碩博士学位論文集"이 있는 데 책자구성은 主題索引(分類索引)만 収録하고 있어 検索은 分類코드로 접근해야 한다.

III. 農学系列 索引語分析

1. 韓国語 意味語 分析

1.1 韓国語 索引語의 構造的 性質

색인어는 索引의 표목으로 채택되는 개개 단어를 이름한다. 한국어 단어는 形態素, 品詞, 複合概念, 意素·語素関係, 名詞의 下位類型에 의해 分析된다.

a. 形態素

依存形態素와 自立形態素로 구분된다.[27] 즉, 임의의 한 단어 "생리적 품종"은 생리·적·품종→自立+依存+自立形態素의 형식으로 분석된다.

b. 品詞

통상 9품사로 分類되는데, 索引語에는 영어의 경우 名詞, 動名詞, 形容詞, 分詞가 선택되며, 한국어는 "주제명표목표[28]", "농업영어집"[29], "한글 농업용어집"[30]에서 보면 거의가 명사형 명사이며, 형용사형이 간혹 複合語의 前置語로 온다.

c. 複合概念

한국어 단어는 단일개념과 두개 이상의 단일개념이 합해진 복합개념을 갖는다. 예를 들면, "균→名 : 자립, 시들음·병→名+名 : 자립+자립"과 같다(名→명사, 자립→자립형태소).

d. 意素·語素관계

단어는 意素와 語素로 이루어진다. 語素에 대한 意素関係는 1) 単意型 : 칼슘-K, 2) 多意型 : 酸性·山城-산성, 3) 同意型 : 抵抗性-저항성·내성과 같이 3가지로 구분한다.

e. 名詞의 下位類型 分類

名詞의 下位類型에 따라서 단어를 분류하면 事物語, 観念語, 動作語로 크게 나눌 수 있으며, 이 안에서 다시 의미상으로 細分하면 각개 範疇를 형성하는 最高上位語를 상대적으로 類推할 수 있다.

예) 事物語 — 集合語 : 有機物名, 無機物名, 生物名, 化合物名
事物語 — 各個語 : 有機物名, 無機物名, 生物名, 化合物名
事物語 — 固有語 : 地名, 人名, 機関·施設名
観念語 — 性質語 : 生物的, 化学的, 物理的, 抽象的 性質
観念語 — 抽象語 : 表徴, 状態, 理論, 尺度, 構想 시간, 表色
動作語 ——— 製作, 反応, 加工, 栽培, 飼養, 防除, 治療, 構築, 경영, 해부, 유전

27) 이길록, 国語文法研究, 서울; 일신사, 1981.
28) 이재철, 주제명표목표, 서울; 연세대학교, 1961.
29) 농촌진흥청, Op. Cit.
30) 농촌진흥청, Op. Cit.

이와 같은 最高上位語는 다시 細分되는 準最高上位語를 갖는다. 예를 들면 有機物名은 농산물, 세포, 조직 등으로 분류된다.

2. 農学索引語의 특성

농학분야의 색인어는 농학을 구성하고 있는 学問分野에 따라 構造 및 특성에서 차이가 있다. 농학색인어는 화학계열, 생물계열, 공학계열, 경제계열의 용어로 분리할 수 있다.

2.1. 화학계열 색인어

化学系列, 用語들은 元素, 化合物名의 事物語, 反応式, 加工 등의 動作語, 이론, 성질, 상태 등의 観念語로 나누인다. 이들 중 특징을 갖는 化合物名은 색인어 중 34%를 차지하며[31], 다른 사물어, 동작어, 관념어와는 달리 화학명, 분자구조식으로 표현이 되고, 또 一般名을 갖는 것들이 있어서 통상 3개 명칭으로 대변이 된다.

화합물명 중 無機化合物은 週期律表上에 기록되어 있는 103가지 元素들이 결합하여 각종 화합물을 구성하고, 유기화합물은 탄소(C), 산소(O), 수소(H), 질소(N), 유황(S), 인(P), 염소(Cl), 브롬(Br) 등의 원소들이 主軸을 이루는 結合을 통해 거의 無限大에 가까운 화합물을 낳고 있다. 따라서 이들의 索引方法이 문제로 대두된다.

화합물명을 索引하는데는 CAS의 연간 색인에서 채용하고 있는 名命法에 의한 방법, Wiswesser의 표기법, Gremas코드법 등 코드에 의한 방법, 원소결합 관련에 의한 원소결합관련법 등 여러 가지가 있다.[32] 이들 중 명명법은 일반명 또는 化学名을 그대로 사용하는 것으로(벤젠, 페놀 등등) 여타의 索引語와 同一하게 취급되어 질 수 있으며, 한가지 문제가 야기된다면 한 색인어에 너무 많은 자소가 연결되어 식별하기 곤란한 경우를 誘発한다. 코드에 의한 방법은[33),34)] 한 마디로 말해서 化学用語 및 主題構成을 分析하여 파셑트적 분류를 하고, 索引時에 분류코드를 서로 合成하여 사용하는 방법을 말한다. 이러한 방법은 과거 20~30년간 欧美諸国에서 刻苦의 노력이 있어 오늘날 많은 종류의 코드법이 개발되어 사용되고 있다.

2.2. 생물계열 索引語

생물계열의 용어는 植物名, 動物名, 微生物名, 組織名인 事物語와 病名, 病徴의 観念語, 飼養, 재배, 번식, 유전 등의 동작어가 존재한다. 이들 중 특징을 갖는 동식물명, 미생물명, 병명을 분석한다.

(1) 동식물명 : 동식물명은 蟲, 나방, 벌레, 나무, 풀, 꽃 등과 같이 자립적 접미어와 결합되는 種名, 이러한 접미어를 갖지 않는 독립적인 種名과 같은 種内의 품종을 표시하는 亞種名으로 분석된다.

예) 1) 種名
- ① 나무 : 밤나무, 소나무, 향나무
- ② 나방 : 복숭아심식나방, 뽕나방
- ③ 예외 : 벼, 보리, 소, 말

2) 亞種名 : 亞種名은 種名을 포함하는 것과 포함하지 않는 것이 있다.
- ① 포함하는 것 : 통일벼, 한국소
- ② 포함하지 않는 것 : 메리노종(면양)

(2) 微生物名 : 微生物名에는 細菌, 바이러스, 真菌, 粘菌病名 있는 바, 病名 옆에 "—균", "—원균"을 붙여 벼·흰꽃 잎마름·병·균(병원균) 또는 学名대로 슈도모나스·브라시케(Pseudomonas, brassicae)로 쓴다.

(3) 病名 : 病名은 複合語로서 前置한 위치에 寄主名, 중간에 表徴내지 특징을 나타내는 形態

---

31) Robert T. Bottle, "Title Indexes as Alerting Services in Chemical and Life Sciences" JASIS, Vol. 21, No1 (1970), pp. 16－22.

32) 金子英昭, "JICST科学技術用語 シソーラス－有機化合物－" 情報管理, Vol. 23, No11(1981), pp. 969－979.

33) Robert Fugman, "The Supply of Information on Chemical Reactions in the IDC system", Information Processing & Management, Vol. 15(1979), pp. 303－323.

34) 이 정일, "化学文献의 電子計算化에 있어서 化学構造 表示法의 役割", 情報管理研究, Vol. 6 (1973), pp. 137－143.

素, 後置에 病이 온다.

예) 포도나무·새눈무늬 · 병

生物用語들은 한자, 영어, 한글이 혼용되며, 病名, 微生物名은 複合語만이 출현하여서 단어길이가 긴 것을 특징으로 한다.

2.3. 工学系列 索引語

工学系列에는 機械, 材料, 装備, 生産品, 動力源의 事物語와 工法, 数式, 状態의 観念語, 그리고 加工(製造), 製作의 動作語가 있다. 여기서 数学主題分野에서는 数式도 索引語로 사용하는 例로 보아[35] 工学系列에 索引語의 一員으로 数式을 포함시킨다.

예) 성재함수율 : $\mu \frac{W_g - W_o}{W_o} \times 100$

2.4. 経済系列 索引語

経済系列用語는 생산품, 기관명, 시설 등의 事物語, 관리, 경영, 계획 등의 動作語, 시장, 정책 등의 観念語로 나누어진다.

3. 英語 의미어 분석

농학계 한국어 문헌에 나타나는 영어단어의 출현 현상을 보면 표제와 표, 도표의 제목은 문장형으로 나타나고 내용에서는 단어형으로 나타난다. 이들 영어단어의 종류를 식별하면 문장형중의 관사, 조사 등을 제외하면 대부분이 명사와 명사형이라고 할 수 있다. 또한 명사를 유형별로 분류하면 생물명, 화합물명, 원소명, 이론명, 인명, 특수기호(도량형 기호) 등이었다.

영어단어중 특수기호, 조사, 관사, 부사를 제외한 명사, 명사형의 단일어와 복합어는 색인용어 정의중 일차로 의미어에 전부 포함한다고 볼 수 있다.

Ⅳ. 농학문헌의 어휘통계

어휘통계의 목적은 한국어색인어의 형태를 통제하고 색인어휘집 형성시 성질항종류의 유추와 분류코드 배정에 응용할 수 있으며, 자동색인 작성시에 입력단어들의 길이를 알 수 있다.

어휘통계는 간단하게 초록, 표제, 결론 등 문헌의 특정위치에서 약식으로 계산할 수 있으나 본고에서는 농학연구자의 이용율이 높은 학회지상의 기사를 그 대상으로 삼았다.

6개 주제분야의 학회지에서 5개씩 30개 기사를 표집하여, 먼저 출현단어들의 품사적 양상을 파악하기 위하여 총출현단어수, 총출현단어종류, 품사별단어출현수를 집계한 결과(표 2참조) 명사가 전체의 75%를 차지하였으며, 나머지 동사, 형용사 등이 25%를 나누어 가졌다.

한편 出現回数를 보면 명사가 36%, 助詞가 32%이고, 나머지 品詞가 32%를 이루었다. 따라서 이용자가 사용한 단어 중 36%가 일차 索引語対象에 오른다고 볼 수 있다.

한편 명사들 중에서 言語学的인 結合状態를 알기 위해서 명사의 複合概念状態를 集計하였는바, 대부분(98.8%)이 3複合概念 이내에 드는것을 알 수 있었다.

그리고 명사의 단어길이를 比較·検討 하였는데 95%가 2音節에서 6音節 사이에 있는 것으로 集計하였다. 따라서 앞의 2개 범주에서 벗어나는 명사들은 단어길이의 길고 짧음, 또는 단어복합개념이 복잡한 理由로 索引語上에서 淘汰되어 질 수 있다. 또한 명사의 類型(事物, 動作, 観念)別 出現様相과 식물병리학분야의 最高上位語別 出現統計를 내었는데, 여기서 보면 名詞中 事物語가 40%, 観念語가 37.5%, 動作語가 22.5%이었고, 最高上位語別로 보면 생물, 화합물, 유기

표 2. 출현단어집계

	종 류		빈 도	
	갯 수	%	횟 수	%
합 계	3429	100%	21,056	100%
명 사	2792	75%	7601	36%
동 사	432	12%	2820	13%
형용사	180	5%	930	4%
조 사	25	1%	6800	32%
기 타	250	7%	2905	15%

35) Winfried Gödert, "Subject Headings for Mathematical Literature", J. of Doc., Vol. 36, No.1(1980), PP. 11-23.

물, 무기물, 지명, 인명, 상태·모양, 표징, 구상, 위치, 표색, 재배, 방제, 유전, 해부 등이 집중적으로 출현한 것을 알 수 있어 디소러스(어휘집)作成에 참조할 수 있다.

영어단어의 출현비율은 전체 한국어단어 3,429개에 비해 536개였으며 출현빈도는 1,697회였다. 이중 도량형기호가 45개 종류에 690회 출현하였다. 그러므로 영어의미어는 491개 종류로 출현횟수는 1,007회가 되는 셈이다.

V. 자동화실험

1. 색인어의 자동 발췌

색인어의 자동 발췌를 위하여 출현현상을 이론적으로 설명할 수 있는 확률색인 이론중 의미어와 무의미어 분류가 가능한 Harter의 two-poisson분포 함수에서 유도한 Z치를 가지고 실험하였다. 2-poisson함수에서 임의단어 wi의 $\lambda_1\lambda_2$가 정해지면 $Z=\frac{\lambda_1-\lambda_2}{\sqrt{\lambda_1+\lambda_2}}$로 구해진다.

본 실험방법은 먼저 통제집단으로 기사 30개를 6개 주제분야별로 5개씩 유층무선표집하여 문헌에 나타난 단어들을[36] 문헌별 출현횟수와 함께 컴퓨터에 입력시켜 Z치를 구하였다. 출력된 단어와 Z치를 비교·검토한 결과, 의미어가 70%이상 포함되어 있는 Z값은 1과3이었다. 이 값을 식별치로 정하고 다른 식물병리분야의 기사 30개에서 실험집단의 단어 300개를 무선표집하여 각 단어의 Z치를 계산하여 출력한 결과, 적정율, 재현율은 각각 60%와 62%였다. 한편 영어단어도 이와 비슷한 성공율을 얻었다.

2. 색인어휘집(디소러스) 자동형성

Salton의 방법론중 제1방법은 실제 실험결과 성공율이 매우 낮았으며,[37] 제2방법은 필자가 식물병리분야의 표 3과 같이 출현한 통계치에서

표 3. 식물병리분야 출현단어유형

		최고상위어 및 준최고상위어		%
사물어	집합어 각개어	생물명	작물* 과수* 채소* 꽃* 세균* 진균* 바이러스*	24%
		화합물명	농약* 양분*	
관념어	추상어	표징	병징*	12%
		상태	병* 생리*	
동작어		방제	*	18%
		재배	*	
		해부	*	
		유전	*	

표 4. 계층구별표

(→: 하위어로 감)

	항 목	구 별	점 수
1	복합개념정도	小→多	
2	단어의 길이정도	短→長	
3	관념어, 사물어구분	관념→사물	
4	계층구분어(계, 과 등)	有→無	
5	性質項語의 포함유무	有→無	
	합 계		

性質項을 16개 선정하여 (표 3의 *표항목) 표 4의 방법으로 적절히 점수(1, 3, 5, 7)를 배정하여 실험한 결과, 주제를 아는 힘이 클수록 형성율이 높았으며 그 반대의 경우는 극히 낮았다. 그러므로 본 실험에서는 1방법과 2방법을 혼합시켜 반자동식으로 어휘집형성을 시도하였다.

그 순서는 1) 성질항을 定한다. 2) 성질항 점수배당 규칙을 작성한다. 3) 의미어들의 성질항을 결정한다. ((1)같은 문장에 성질항어와 함께 출현할때 그 성질항에 2점, (2) 같은 문단에 성질항어와 함께 출현할때 그항에 1점 (3) 어느 성질항어와도 출현하지 않을 때 Clustering을 하여 나타나는 관련어의 성질항에 1점을 준다.) 4) 단어대 성질항의 행렬을 작성하며 동일단어

36) 띄어쓰기의 불분명으로 인하여 명사, 명사형 복합어들의 구분은 무조건 붙여쓴 음절을 한단어로 취급하였으며, 조사는 분리하여 독립된 한 단어로 처리하였고, 동사는 과거, 현재, 수동형(~되다), 능동형(~하다)의 네가지 형태로 집약하였다.

37) 정문정, 「Clustering for Information Retrieval System」. 한국과학원석사학위논문집. 1975.

들의 점수를 아울러 합산한다. 5) 4의 행렬값을 2번의 점수배당 方法으로 수정보완한다. 6) Salton의 제1방법으로 정리한다. 7) 표 5의 규칙에 어긋나는 것들은 동의·계층관계에서 친화관계로 수정한다.

〈실험수행 및 결과〉

실험집단은 15개 문헌에서 의미어 50개를 발췌하여 실험하였다. 수행과정과 결과를 의미어 8개로 예를들어 설명하면 표 6과 같이, 같은 문장 문단에서 출현한 성질항의 괄호안에 숫자를 주었으며 사선아래의 숫자는 5번 과정을 거친후 배당된 것이다. 여기서 *표한 성질항은 표 3에 없던 차하급 하위항목을 등장시켰는데, 이는 실험문헌을 조사한 결과 어떤 성질항목들은 의미어와 함께 출현하지 않았기 때문에 보완한 것이다. 점수주는 방법은 상위어가 높은 점수를 받도록 하였다. 표 6과 같이 정렬된 행렬을 Salton의 제1방법으로 어휘집을 형성시켜보니, 형성율을 수치로 나타내기는 어려우나 필자가 디소러스를 작성해본 경험에 의하면 기초작업으로서의 역할은 충분할 것으로 사료되었다.

표 5. 단어유형의 계층관계

(→ : 하위어로 감)

단어유형	예
1) 관념어→동작 • 관념어 ↘ 사물어	1) 생리→동화작용 • 물질대사 ↘ 영양소
2) 사물어→사물어 ↘ 관념 • 동작어	2) 식물→작물 ↘ 식물생태 • 벼재배
3) 동작어→동작 • 관념어 ↘ 사물어	3) 육종→식물육종 • 육종방법 ↘ 양

Ⅵ. 색인어휘의 시스템적 통제

Ⅲ, Ⅳ, Ⅴ장에서 농학계열 색인어휘의 분석과 통계 그리고 자동화에 대한 적응도를 고찰하였다. 그 결과 일련의 색인어휘에 대한 속성을 파악하게되었는데 분명히 색인어는 문헌내의 자료소(단어)에서 발산되며 변형함수[38]에 의해 색인어다워져야 함을 알았다. 따라서 본장에서는 앞서의 연구결과에 의한 경제적이고 효율적인색인어휘 시스팀을 제시한다.

1. 색인어 형태 통제

1) 한 색인어는 最少自立形態素를 1~3개로 제한한다(複合概念이 너무 많으면 단어의 表意性이 模糊해진다).[39]

표 6. 단어 대 성질항 행렬표

	세 균	진균병*	살균제	균사*	곡류작물*	……
풋마름병균	(2)/ 2					
흑색뿌리썩음병		2		(3)		
Dithane			(3)/ 2			
종자소독제			(1)/ 4			
자낭각+				(2)/ 2		
자낭포자+				(2)/ 2		
분생포자+				(3)/ 2		
맥류					(2)/ 4	

(* : 표 3에 없던 항목, + : Clustering한 의미어,
수정점수종류: 2, 4, 6)

38) Landry, op. cit..

39) Kevin P. Jones, "Towards a Theory of Indexing", J. of Doc. , Vol. 32, No2 (1976), pp. 118-125.

〈예〉 肉汁·감자·한천·배양기→혼합배양기
〈예외〉 병명, 생물명, 화학명, 공식은 위의 제한을 받지 않는다.

2) 倒置型의 索引語는 사용하지 않는다(어휘에 特定性을 주기 위해서이다).
예) 식물·해부 ↛ 해부, 식물

3) 略字는 索引語로 사용하지 않는다(단어의 길이가 너무 짧으면 意味表現이 불확실해지는 경우가 많다[40]).
〈예〉 산경연→산업경제기술연구원, 농진청→ 농촌진흥청.
〈예외〉 화학명 중 화합물명은 上記의 統制를 받지 않으며, 英語名의 機関名 및 기구의 略名은 上記의 統制를 받지 않는다.

4) 한 문헌에 配当되는 색인어는 10개 이하로 制限한다.
〈예외〉 기본색인부에서는 색인어를 15개 이하로 제한한다.

5) 한 索引語는 10음절 이하로 제한한다(단어의 길이가 너무 길면 그 나타내는 뜻이 模糊해진다). 病名, 病原菌名은 제외한다.

6) 基本索引語를 単一語로 限定한다.
〈예외〉 依存自立素와 形容詞는 종속되는 명사에 붙여 한 単一語取扱을 해 준다.
〈예〉 세포 · 질→세포질, 검은 · 나비→검은나비

2. 색인어휘집(디소러스) 作成

2.1. 索引語의 形成

모든 索引語는 한국어로 쓰고, 우리말이 없는 경우에는 한글로 飜字하여 쓴다. 또한 필요하다면 ( )안에 原語名을 附記한다.

2.1.1. 化学系列 索引語

1) 事物語中 元素名과 化合物名의 略語名, 一般名, 짧은 化学名(로마자 20자 이하인 것), 짧은 분자구조식(화학명으로 로마자 20자 이하인 것)은 上記의 形成準則에 따른다.
〈예〉 P→인, EPN제→이피엔제, Methane→메탄, CO_2→탄산가스,

2) 긴 화학명과 긴 분자구조식(화학명으로 로마자 20자 이상되는 것)은 화학명을 한글로 飜字하여 쓰되 語彙集内에서 별도로 로마자 表記의 英語名으로부터 참조를 내준다. 만약에 一般名이 있으면 一般名으로 代替시키며, 이 관계를 語彙集内에서 참조로써 연결시킨다.

〈예〉 CH_3O $P-CH-CCl_3$→ O, O-dimethylhydrogenphosphate→ 디프테렉스 (Dipterex)
CH_3O

2.1.2. 生物系列 索引語

1) 생물명 : 생물명 중 属名, 種名, 亜種名, 変種名은 一般名을 한국어로 쓰고, 그 学名이 出現할 경우는 学名을 ( )안에 附記하고 学名이 出現안할 경우에는 学名 중 属名을( ) 에 併記한다.

〈예〉 소나무(Pinus, Strobus)
↑
학명이 출현한 예.
곰솔(Pinus)
↑
학명이 출현안한 예.
또한 門, 綱, 目, 科, 属, 種의 분류단위 명칭 중 種을 제외한 것은 단위명칭인 門, 綱, 目, 科, 属을 붙여서 사용하고, 단위명칭이 붙은 라틴어 学名을 ( )속에 附記한다.
〈예〉 버드나무과(Salicaceae)
여기서 微生物名은 한글 飜字上의 混乱을 피하기 위해 라틴어학명에서 한글飜字名으로 語彙集内에서 別途로 참조를 내준다.

2) 그외 事物語와 動作語, 観念語는 索引語形成準則을 따른다.

40) Kevin P. Jones, Opocit.

2.1.3 工学 및 経済系列 索引語.

1) 数式(公式)은 数式名을 索引語로 삼고 함수는 사용하지 않는다.
2) 그외 索引語는 索引語形成準則에 따른다.

2.2. 索引語関係決定

색인어의 統制는 근본적으로 색인의 一貫性維持와 検索再現率을 높이기 위한 手段이다. 현재 디소러스는 백수십종이 있으며, 農学関係만 11種이 있다[41]

본 논문에서는 디소러스의 관계표시 종류 및 기호는 AGROVOC과 TEST(Thesaurus of Engineering and Scientific Terms)의 방법을 따른다.

(1) 対等関係

対等関係에서 取扱되는 用語들의 관계는 同意語와 準同意語関係로 分離하여 생각할 수 있다.

1) 同意語関係

同意語는 商品名, 사투리, 略語, 異語素語 유행어에서 파생하는 語素는 다르되, 意素는 같은 対等関係를 統制해 주는 것을 뜻한다.

①異語素語

〈예〉 耐性, USE, 抵抗性

②商品名

〈예〉 디, 피, 씨제(D. P. C. 제)UF 카라단Ⓡ (karathane)

본 설계에서도 原名을 사용하되, 原名이 길거나(화학명), 분자구조식으로 나왔을 경우에만 상품명으로 사용한다.

③ 약어

2) 準同意語

準同意語概念에 포함되는 用語関係는 反意語特定概念語가 있다.

① 反意語

〈예〉 感受性, USE, 抵抗性,

②特定概念語 : 일정한 概念 아래 나열되는 用語들을 上位概念으로 묶어주는 것을 말한다〈예1)〉. 또한 同一概念中 代表語를 선정하여 디스크립터로 한다〈예 2〉.

〈예 1〉 괴저, UF, 피층괴저, 유조직괴저, 반점성괴저

〈예 2〉 종양, UF, 암종, 우점, 염류

(2) 階層関係

索引의 깊이에 관련되는 문제로서 피드백을 하며, 上下探索을 가능케 한다.

1) 後属関係 : 식물, 미생물, 동물명들이 이에 속한다.

예) 곤충 NT 나비

2) 部分/全体関係

〈예〉 서울 BT 한국

複合語일 때의 名詞下位類型은 항상 後置語에 後属한다.

3) 親和関係.

概念的으로는 매우 근접해 있지만 階層的으로는 다른 用語들을 서로 이어주는 관계로

① 사물/부분 : 세균 RT 전염병
② 사물/성질 : 알코올 RT 휘발, 방향
③ 사물/과정 : 병원균 RT 전염
④ 사물/응용 : 향나무 RT 중간기주.
⑤ 성질/과정 : 신맛 RT 초산 발효와 같은 関係를 맺어준다.

이상과 같이 디소러스作成의 細部規則을 서술하였다.

Ⅶ. 結論 및 提言

農学係 韓国語, 英語, 日語 資料의 混合出力을 위하여 우선적으로 韓国語문헌상에 出現하는 한국어와 영어 意味語들의 현상을 文法的, 통계적으로 分析하였다. 일차적으로 의미어상에 오른 단어는 품사별로 명사, 명사형임을 알았고 통계적으로는 복합개념이 3개이내, 단어 길이는 2음절에서 6음절사이에 대부분(95% 이상)소재하고 있는 것을 알았다.

41) Maxine McCatterty, Thesauri & the Thesauri Construction, Aslib bibliography №7, 1977, pp. 114－118.

Harter의 이론으로 의미어자동발췌를 실험하여 60%의 성공률을 얻었으며, 각의미어간의 어휘집형성을 salton의 이론을 수정하여 실험한 결과 타당성을 발견하였으며 위 현상에서 요약된 합리점을 모아 경제적인 측면과 효율성을 중요시한 어휘통제시스팀을 작성하였다.

궁극적으로 일정한 法則아래 색인어들의 자동발췌가 가능해지며 필요에 따라 어휘집의 형성이 이루어지는 상태로 발전하려면 복합어와 단일어의 개념수 조정, 성질항의 가지수 조정, 문헌군수의 조정에 관한 연구와 새로운 이론의 출현이 기대된다.

〈 참 고 문 헌 〉

1. 김 민수, 国語意味論, 서울 : 일조각, 1981.
2. 사공철, 「정보검색에 있어서의 Thesaurus 도입에 관한 기초연구」, 석사학위논문, 연세대학교 산업대학원, 1974.
3. 최성진, 情報学原論, 서울 : 아세아문화사, 1976.
4. 森岡祐=, "BIOSIS(生物)," 情報管理, Vol. 23, No. 8 (1980). pp. 721－736.
5. 中山和彦, "CAB(農学)", 情報管理, Vol. 23, No. 9(1980). pp. 836－850
6. Alexander, M. J., Information System Analysis : Theory and Applications, Kingsport: Kingsport press, 1974
7. Borko, H., "Design of Information Systems and and Services", American Documentation Institute－Annual Review of Information Science and technology, Vol. 2, (1967), pp. 35－61.
8. Coates, E. J., "Some Properties of Relationships in the Structure of Indexing Language", Progress in Documentation, Vol. 29, No4(1973)
9. Heaps. H. S., Information Retrieval: Conceptual & Theoretical Aspects, New York, ACADEMIC Press, 1978.
10. Jones, K. S., "Some thesauric history", Aslib Proceedings, Vol. 24, No. 7, 1972. pp. 400－411.
12. Kraft, D. H., "Evaluation of Information Retrieval Systems: A Decision Theory Approach", JASIS, Vol 29. No1, (1978) pp. 31－40.
12. Larson, J. R., "Comparison of Printed Bibliographic Descriptions Distributed by Biosis. CAS. and Ei ,", JASIS, Vol., 27, No. 1(1976)
13. Maron, M. E., "Associative Search Techniques Versus Probabilistic RetrievalModels", JASIS V L 33. No5(1982). pp 308－310.
14. Robertson, S. E., "The Probability Ranking Principle in IR ", J. of DOC. , Vol. 33, No4. (1977), pp. 294－304.
15. Salton, G., "A Theory of Term Importance in Automatic Text Analysis", JASIS Vol. 26. No. 1(1975), pp33－44.
16. Salton, G., "The Measurement of Term Importance in Automatic Indexing", JASIS Vol. 32, No. No. 3(1981). pp. 175－186.
17. Stile. H. E., "The Association factor in Information Retrieval, "JACM Vol. 8, No2 (1961) pp. 271－279.
18. Van Lijsbergen, C. J., "A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval", J. of Doc., Vol. 33, No. 2(1977) pp. 106－119.
19. Van Lijsbergen, C. J., "An Evaluation of Feedback in Document in Retrieval Using Co-occurreuce data", Joof Doc., Vd. 34, No3(1978) pp. 189－216.