NATURAL LANGUAGE TEXT SEGMENTATION TECHNIQUES APPLIED TO THE AUTOMATIC COMPILATION OF PRINTED SUBJECT INDEXES AND FOR ONLINE DATABASE ACCESS

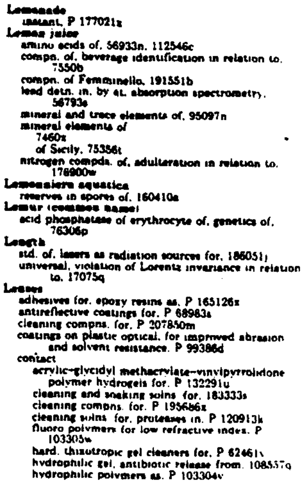
G. Vladutz
Institute for Scientific Information
3501 Market Street, Philadelphia, Pennsylvania 19104 USA

ABSTRACT

The nature of the problem and earlier approaches to the automatic compilation of printed subject indexes are reviewed and illustrated. A simple method is described for the detection of semantically self-contained word phrase segments in title-like texts. The method is based on a predetermined list of acceptable types of nominative syntactic patterns which can be recognized using a small domain-independent dictionary. The transformation of the detected word phrases into subject index records is described. The records are used for the compilation of Key Word Phrase subject indexes (KWPSI). The method has been successfully tested for the fully automatic production of KWPSI-type indexes to titles of scientific publications. The usage of KWPSI-type display formats for the enhanced online access to databases is also discussed.

1. The problem of automatic compilation of subject indexes

Printed subject indexes (SI), such as back-of-the-book indexes and indexes to periodicals and abstracts journals remain important as the most common tools for information retrieval. Traditionally SI are compiled from subject descriptions produced for this purpose by human indexers. Such subject descriptions are usually nominalized sentences in which the word order is chosen to emphasize as rheme one of the objects participating in the description; the corresponding word or word phrase is placed at the beginning of the nominative construction. Furthermore, the nominalized sentence is rendered in a specially transformed ('articulated') way involving the separated by commas display of component word phrases together with the dominating prepositions; e.g. the sentence 'In lemon juice lead (is) determined by atomic absorption spectrometry' becomes 'LEMON JUICE, lead determination in, by atomic absorption spectroscopy.' Such rendering enhances the speedy understanding of the descriptions when browsing the index. At the same time it creates for the subject description a lineary ordered sequence of focuses which can be used for the hierarchical multilevel grouping of related sets of descriptions. The main focus (rheme) serves for the grouping of descriptions under a corresponding subject heading, the secondary focuses make possible the further subdivision of such group by subheadings. This is illustrated on the SI fragment to "Chemical Abstracts" shown on figure 1.



Lemonade
instant, P 177021x
Lemon juice
amino acids of, 56933n, 112546c
compn. of, beverage identification in relation to, 7550b
compn. of Femminello, 191551b
lead detn. in, by at. absorption spectrometry, 56793e
mineral and trace elements of, 95097n
mineral elements of
7460z
of Sicily, 75356t
nitrogen compds. of, adulteration in relation to, 176900w
Lemonniera aquatica
reserves in spores of, 160410a
Lemur (common name)
acid phosphatase of erythrocyte of, genetics of, 76306p
Length
std. of, lasers as radiation sources for, 186051j
universal, violation of Lorentz invariance in relation to, 17075q
Lenses
adhesives for, epoxy resins as, P 165126x
antireflective coatings for, P 68983s
cleaning compns. for, P 207850m
coatings on plastic optical, for improved abrasion and solvent resistance, P 99386d
contact
acrylic-glycidyl methacrylate-vinylpyrrolidone polymer hydrogels for, P 132291u
cleaning and soaking solns. for, 183333s
cleaning compns. for, P 195686z
cleaning solns. for, proteases in, P 120913k
fluoro polymers for low refractive index, P 103305w
hard, thixotropic gel cleaners for, P 62461v
hydrophilic gel, antibiotic release from, 108557q
hydrophilic polymers as, P 103304v

Figure 1
Fragment of a subject index of traditional type to "Chemical Abstracts," compiled from subject descriptions by human indexers. A text processing problem, studied in connection with the compilation of such SI of traditional type, was the automatic transformation of subject descriptions for selecting their different possible rhemes and focuses (Armitage, 1967). An experimental procedure, not yet implemented, takes as input pre-edited subject descriptions (Cohen, 1976).

Since the generation of subject descriptions by human indexers is a very expensive procedure P. Luhn (1959) of IBM has suggested replacing subject descriptions by titles provided by the publication's authors. Using only a 'negative' dictionary of high frequency words excluded from indexing, he designed a procedure for the automatic compilation of listings where fragments of titles are displayed repeatedly for all their indexable words. These words are alphabetized and displayed on the printed page in the central position of a column; their contextual fragments are sorted according to the right-hand side contexts of the index words. Such listings, called Key-Word-in- Context (KWIC) indexes, have been produced and successfully marketed since 1960 as "quick-and-dirty" SI, despite their 'mechanical' appearance which makes them difficult enough to read and browse. A fragment of KWIC index to "Biological Abstracts," featuring titles enriched by additional key words is shown in figure 2.

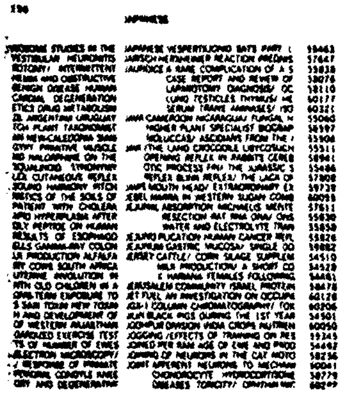
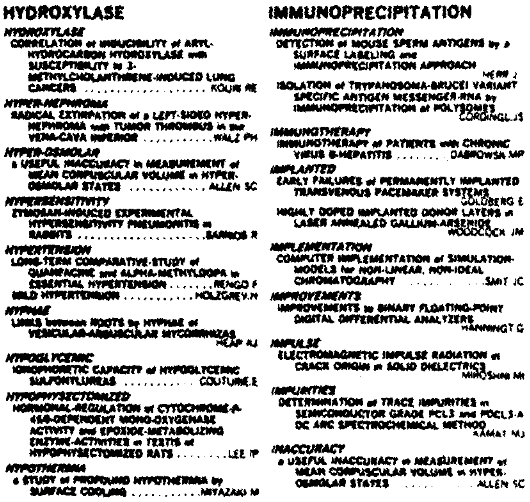


Figure 2.
Fragment of a Key-Word-in-Context (KWIC) type subject index to "Biological Abstracts," automatically compiled from titles of biological publications. The blank spaces replace the repeated occurrences of the key word appearing above.

Another mechanically compiled SI substitute still in current use is based on a similar idea and simply groups together all the titles containing a same indexable word. Such Key-Word-out-of-Context (KWOC) indexes display the full texts of titles under a common heading. Figure 3 shows a KWOC sample generated from title-like subject descriptions at the Institute for Scientific Information. The appearance of KWOC indexes is more acceptabe but their browsing is much hindered by the lack of articulation of the lengthy subject descriptions (titles). Without proper articulation, the recognition of the context immediately relevant to the index word becomes too slow.

In 1966 the Institute for Scientific Information (ISI) introduced a different type of automatically compiled subject index called PERMUTERM Subject Index (PSI) OGarfield, 1976o, which at present is the main type of SI to the Science Citation Index and other similar ISI publications. Two different negative dictionaries are used for producing this SI: a so called "full stop list" of words excluded from becoming headings as well as from being used as subordinate index entries, and a "semi- stop list" of words of little informative value, which are not allowed as headings but are used as index entries along with words found neither in the full-stop nor in the semi-stop lists. In the PSI every word co-occuring with the heading word in some subject description (title) becomes an entry line subordinated to this heading. The format of the PSI is illustrated in figure 4.



HYDROXYLASE

HYDROXYLASE
CORRELATION of INDUCIBILITY of ARYL-HYDROCARBON HYDROXYLASE with SUSCEPTIBILITY to 3-METHYLCHOLANTHRENE-INDUCED LUNG CANCERS .... KOUR RE

HYPER-NEPHROMA
RADICAL EXTIRPATION of a LEFT-SIDED HYPER-NEPHROMA with TUMOR THROMBUS in the VENA-CAVA INFERIOR .... WALZ PH

HYPER-OSMOLAR
a USEFUL INACCURACY in MEASUREMENT of MEAN CORPUSCULAR VOLUME in HYPER-OSMOLAR STATES .... ALLEN SC

HYPERSENSITIVITY
ZYMOSAN-INDUCED EXPERIMENTAL HYPERSENSITIVITY PNEUMONITIS in RABBITS .... BARROS R

HYPERTENSION
LONG-TERM COMPARATIVE-STUDY of GUANFACINE and ALPHA-METHYLDOPA in ESSENTIAL HYPERTENSION .... RENGO F
MILD HYPERTENSION .... HOLZGREV.H

HYPHAE
LINKS between ROOTS by HYPHAE of VESICULAR-ARBUSCULAR MYCORRHIZAS
HEAP AJ

HYPOGLYCEMIC
IONOPHORETIC CAPACITY of HYPOGLYCEMIC SULFONYLUREAS .... COUTURE.E

HYPOPHYSECTOMIZED
HORMONAL-REGULATION of CYTOCHROME-P-450-DEPENDENT MONO-OXYGENASE ACTIVITY and EPOXIDE-METABOLIZING ENZYME-ACTIVITIES in TESTIS of HYPOPHYSECTOMIZED RATS .... LEE IP

HYPOTHERMIA
a STUDY of PROFOUND HYPOTHERMIA by SURFACE COOLING .... MIYAZAKI M

IMMUNOPRECIPITATION

IMMUNOPRECIPITATION
DETECTION of MOUSE SPERM ANTIGENS by a SURFACE LABELING and IMMUNOPRECIPITATION APPROACH
HERR J
ISOLATION of TRYPANOSOMA-BRUCEI VARIANT SPECIFIC ANTIGEN MESSENGER-RNA by IMMUNOPRECIPITATION of POLYSOMES
CORDINGL.JS

IMMUNOTHERAPY
IMMUNOTHERAPY of PATIENTS with CHRONIC VIRUS B-HEPATITIS .... DABROWSK MP

IMPLANTED
EARLY FAILURES of PERMANENTLY IMPLANTED TRANSVENOUS PACEMAKER SYSTEMS
GOLDBERG E
HIGHLY DOPED IMPLANTED DONOR LAYERS in LASER ANNEALED GALLIUM-ARSENIDE
WOODCOCK JM

IMPLEMENTATION
COMPUTER IMPLEMENTATION of SIMULATION-MODELS for NON-LINEAR, NON-IDEAL CHROMATOGRAPHY .... SANT JC

IMPROVEMENTS
IMPROVEMENTS to BINARY FLOATING-POINT DIGITAL DIFFERENTIAL ANALYZERS
HANNINGT G

IMPULSE
ELECTROMAGNETIC IMPULSE RADIATION at CRACK ORIGIN in SOLID DIELECTRICS
MIROSHNI MI

IMPURITIES
DETERMINATION of TRACE IMPURITIES in SEMICONDUCTOR GRADE PCL3 and POCL3-A DC ARC SPECTROCHEMICAL METHOD
KAMAT MJ

INACCURACY
a USEFUL INACCURACY in MEASUREMENT of MEAN CORPUSCULAR VOLUME in HYPER-OSMOLAR STATES .... ALLEN SC

Figure 3.

Fragment of a KWOC index compiled from relatively long subject descriptions. The words priuted in lowercase letters are "stop words," not used as index headings.

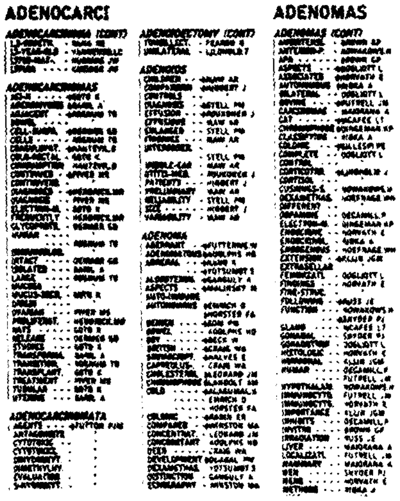

ADENOCARCI
ADENOMAS

Figure 4.
Fragment of a PERMUTERM subject index to the Science Citation Index (Institute for Scientific Information), automatically compiled from titles. The index lines are words co-occuring in titles with the heading word. The arrows indicate the first occurrence under the given heading of a pointer to a given article.

PSI has the unique ability to make possible the easy retrieval of all titles containing any given pair of informative words. This ability is similar to the ability of computerized online search systems to retrieve titles by any boolean combination of search terms. The corresponding PSI ability is available to PSI users who have been instructed about the principles used for compiling it. The naive user is more likely to utilize it as a browsing tool. When doing so, he may be inclined to perceive the subordinate word entries as being the immediate context of the headings. Used as a browsing tool, PSI may deliver relatively high percentage of false drops because of the lack of contextual information. Another shortcoming of the PSI is its relatively high cost due to its significant size which is proportional to the square of the average length of titles. The large number of entries subordinated to headings which are words of relatively high frequency makes the exhaustive scanning of entries under such headings a time consuming procedure.

An important advantage of all the above computer generated indexes over their manually compiled counterparts is the speed and essentially lower cost at which they are made available.

All the above compilation procedures are based exclusively on the most trivial facts concerning the syntaxis and semantics of natural languages. They make use of the fact that texts are built of words, of the existance of words having purely syntactic functions and of the existence of lexical units of very little informative value. A common disadvantage versus the SI of traditional type is that the above procedures fail to provide articulated contexts which would be short enough and structurally simple enough to be easily grasped in the course of browsing.

Certainly this problem can be solved by any systen which can perform the full syntactic analysis of titles or similar kinds of subject descriptions. From the syntactic tree of the title a brief articulated context can be produced for any given word of a title by detecting a subtree of suitable size which includes the given word. However, in the majority of cases the practical conditions of application of index compilation procedures are excluding the usage of full scale syntactic analysis, based on dictionaries containing the required morphological, syntactic and semantic information for all the lexical units of the processed input. For instance, ISI is processing annually for its mutli-disciplinary publications around 700,000 titles ranging in their subject orientation from science and technology to arts and humanities. The effort needed for the creation and maintenance of dictionaries covering several hundred thousands entries with a high ratio of appearance of new words would be excessive. Therefore, the automatic compilation of SI is practially feasible only on the basis of quite simplistic procedures based on "negative" dictionaries involving approximative methods of analysis which yield good results in the majority of cases, but are robust enough not to break down even in difficult cases.

At one end of the range of problems involving natural language processsing are such as question answering which require a high degree of analytic sophistication and are based on a significant amount of domain dependent information formated in bulky lexicons. Such procedures appear to be applicable to texts dealing with rather narrow fields of knowledge in the same way as the high levels of in-depth human expertise are usually limited to specific domains. On the other end of the spectrum are simple problems requiring much less domain dependent information and relatively low levels of "intelligence" (defined as the ability to discuss comprehensive texts from gibberish); the corresponding procedures are usually applicable to wide categories of texts. For reasons explained above, we consider the problems of automatic compilation of subject indexes as belonging to this low end of the spectrum.

2. The automatic compilation of Key-Word-Phrase Subject Indexes

In this framework we developed an automatic procedure for the compilation of a SI based on the detection and usage of word phrases. The earlier stages of development of this Key-Word-Phrase subject index (KWPSI) have been reported elsewhere (Vladutz 1979). The procedure starts by detecting certain types of syntactically self-contained segments of the input text; such segments are expected to be semantically self-contained in view of the assumed well-formedness of the input. The segment detection procedure is based on a relatively short list of acceptable syntactic patterns, formulated in terms of markers attributable by a simple dictionary look-up. The markers are essentially the same as used in (Klein 1963) in the early days of machine translation for automatic grammatical coding of English words. All the words not found in an exlusion dictionary of ∼ 1,500 words are assigned the two markers ADJ and NOUN. All the acceptable syntactic patterns are characterized in the frameworks of a generative grammar constructed for title-type texts. Such texts are described as sequences of segments of acceptable syntactic patterns separated by arbitrary filler segments whose syntactic pattern is different from the acceptable ones. The analysis procedure leading to the detection of acceptable segments was formulated as a reversal of the generative grammer and is performed by a right to left scanning. New acceptable syntactic patterns can easily be incorporated into the generative grammar. It is envisaged to use in the future existing programs for automatically generating analysis programs from any specific variant of the grammar.

The present list of acceptable syntactic patterns includes such patterns where noun phrases are concatenated by the preposition 'OF' and the conjunctions 'AND', 'OR', 'AND/OR', as well as constructions of the type 'NP1, NP2, ... AND NPi'. Since no prepositions other than 'OF' and no conjunctions other than 'AND', 'OR', 'AND/OR' can occur in the acceptable segments the occurrences of other prepositions and conjunctions are used for initial delimitation of acceptable segments, but the detection procedure is not limited to such usage. In particular, a past participle or a group containing adverbs followed by a past participle are excluded from the acceptable segment when preceding an initial delimiter. The segmentation detection is illustrated for three titles in figure 5.

A) SPREADING OF VIRUS INFECTION among WILD BIRDS And MONKEYS during INFLUENZA EPIDEMIC Caused by VICTORIA(3)75 VARIANT OF A(H3N2) VIRUS

B) EXERCISE Induced CHANGES in LEFT-VENTRICULAR Function in Patients with MITRALE-VALVE PROLAPSE

C) DIFFERENTIATION OF MLC-INDUCED KILLER And Suppressor T-CELS by SENSITIVITY to PYRILAMINE

Figure 5.

The detection of acceptable segments is shown for 3 titles. The words with all lowercase letters are prepositions and conjunctions used as initial delimiters. The words with only initial capital letters are "semi-stop" words, excluded from being used as index headings; the underscored by dotted lines "semi-stops" are past participles which become deliminters only when followed by initial delimiters. The resulting multi-word phrases are underscored twice unlike the resulting single word phrases which are underscored once.

The first part of the system's dictionary conjunctions, prepositions, articles, auxiliary verbs and pronouns. This part is completely domain independent. A second part of the dictionary consists of nouns, adjectives, verbs, present and past participles, all of them of little informative value and, therefore, called "semi-stop" words. Such words will not be allowed later to become SI headings. The semi-stop part of the dictionary is somewhat domain-dependent and has to be atuned for different broad fields of knowledge such as science and technology, social sciences or arts and humanities.

The second logical step in the SI compilation involves the transformation of acceptable segments into index records consisting of an informative word (not found in the system's dictionary) displayed as heading line and of an index line providing some relevant context for the heading word. Each multi-word segment generates as many index records as many informative words it contains. The right-hand side of the segment following the heading word is placed at the beginning of the index line to serve as its immediate context and is followed through a semicolon by the segment's left-hand side. When both sides are non-empty, an articulation of the index line is so achieved. In the case of a single word segment an "expension" procedure is performed during index record generation. It starts by placing at the beginning of the index line a fragment of the title consisting of the filler portion following the heading word and of the next acceptable segment, if any; this initial portion of the index line is followed by a semicolon after which follows the preceding acceptable segment, followed finally by the filler portion separating in the title this preceding segment from the heading word. The index record generation is illustrated in figure 6.

The final "enrichment" phase of the index record generation involves the additional display (in parenthesis) of the unused segments of the processed title.

*SPREADING
of VIRUS INFECTION

*VIRUS
INFECTION; SPREADING of *

INFECTION
SPREADING OF VIRUS *

*WILD
BIRDS and MONKEYS

*BIRDS
and MONKEYS; WILD *

*MONKEYS
WILD BIRDS and *

*INFLUENZA
EPIDEMIC

*EPIDEMIC
INFLUENZA *

*SENSITIVITY
to PYRILAMINE; DIFFERENTIATION of
MLC-INDUCED KILLER SUPRESSOR T-CELLS by *

*PYRILAMINE
SENSITIVITY to *

Figure 6.
The transformation of Key-Word-Phrases into subject headings and subject entries is illustrated for the first two segments of the title A, Figure 6. The last two examples show how single word segments (from Title C) are expanded to include the preceding and following them segements.

As a result of this stage the informational value of the finally generated index record is almost equivalent to the information content of the initial full title. The entire process ultimately boils down to the the reshuffling of some component segments of the initial title. The enrichment stage of index record generation is illustrated on figure 7.

The index records are alphabetized firstly by heading words and secondly by index lines with the exclusion from alphabetization of prepositions and conjunctions if they occur at the beginning of index lines. During the photocomposition different parts of the index line are set using different fonts. If in the original title the initial part of the index line follows the head word immediately this part is set in bold face italics, i.e. in the same font as the heading. The "inverted" part following the semicolon is set in light face roman letters. Finally the enrichment part of the index line, included in parens is always displayed in light-face italics. As a result the

```
*SPREADING
   of VIRUS INFECTION (WILD BIRDS and MONKEYS;
   INFLUENZA EPIDEMIC; VICTORIA(#)75 VARIANT
   of A(H3N2) VIRUS)

*VIRUS
   INFECTION; SPREADING OF * (WILD BIRDS and
   MONKEYS; INFLUENZA EPIDEMIC; VICTORIA(3)75
   VARIANT of A(H3N2) VIRUS)

*BIRDS
   and MONKEYS; WILD * (SPREADING of VIRUS
   INFECTIONS; INFLUENZA EPIDEMIC, EPIDEMIC;
   VICTORIA(3)75 VARIANT of A(H3N2) VIRUS)

*INFLUENZA
   EPIDEMIC (SPREADING of VIRUS INFECTIONS;
   VICTORIA(3)75 VARIANT of A(H3N2) VIRUS)

*PYRILAMINE
   SENSITIVITY to * (DIFFERENTIATION of MLC-
   INDUCED KILLER and SUPPRESSOR T-CELS)
```

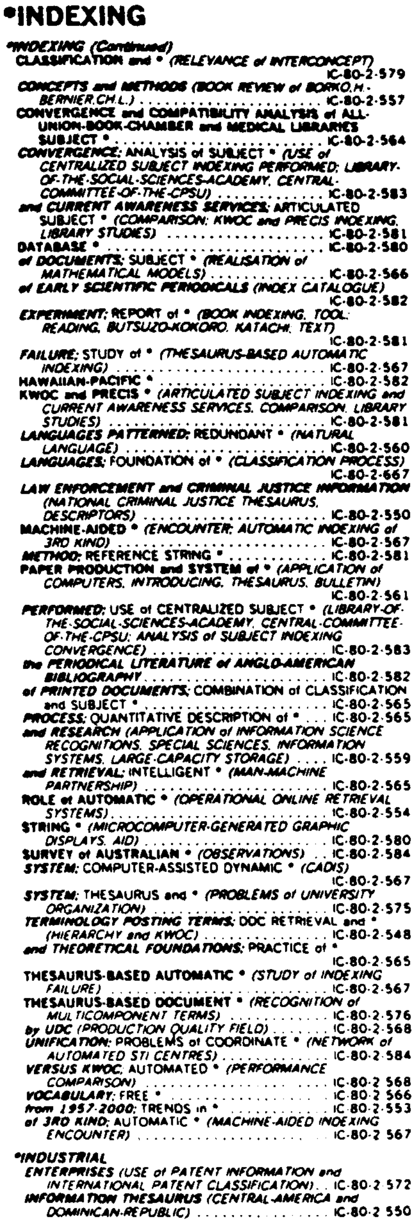
Figure 7
The enrichment of the subject entries by the display (in parenthesis) of the unused by them segments of the same title, illustrated for some of the entries of Figure 6.

immediately relevant context of the head word is displayed in bold face in order to facilitate its rapid grasping when browsing. Details of the appearance and structure of KWPSI are exemplified in figure 8 on a sample compiled for titles of publications dealing with librarianship and information science. The general appearance of KWPSI is close enough to the appearance of SI of traditional type.

For purposes of transportability the KWPSI system is programmed in ANSA COBOL. It includes two modules: the index generation module and the sorting and reformatting module. On an IBM 370 system index records are generated for titles of scholarly papers at a speed of ~ 70,000 titles/hour. The resulting total size of the index is of the same order as the size of KWOC indexes and compares favorably with the size of the PSI index.

The analysis of the rates and nature of failures of the segment detection algorithm shows that in 96% of cases the generated segments are fully acceptable as valuable index entries. In 2% of cases some important information is lost as a result of the elimination of prepositions, as in case of expressions of 'wood to wood' type. The rest of failures results in somehow awkward segments which are not completely semantically self-contained. Even in such cases the index entries retain some informative value. Around half of the failures can be eliminated by additions to the system's dictionary, especially by the inclusion of more verbs and past participles. Not counted as failures are the 5% of cases when the length of the detected segments is excessive; such segments can include the whole title.

The extent of tuning required for the application of the system in a new area of knowledge depends mainly upon the extent of figure 8. deviations from the normal structure of natural language texts occurring in the new file. As a matter of fact all kinds of scholarly titles contain such deviations, as for instance portions of normal text included in parentheses or occurrences of mathematical or chemical symbols. We found only one case when the required tuning effort was significant, namely the case of titles from the domain of arts and humanities. ISI's "Arts and Humanities Citation

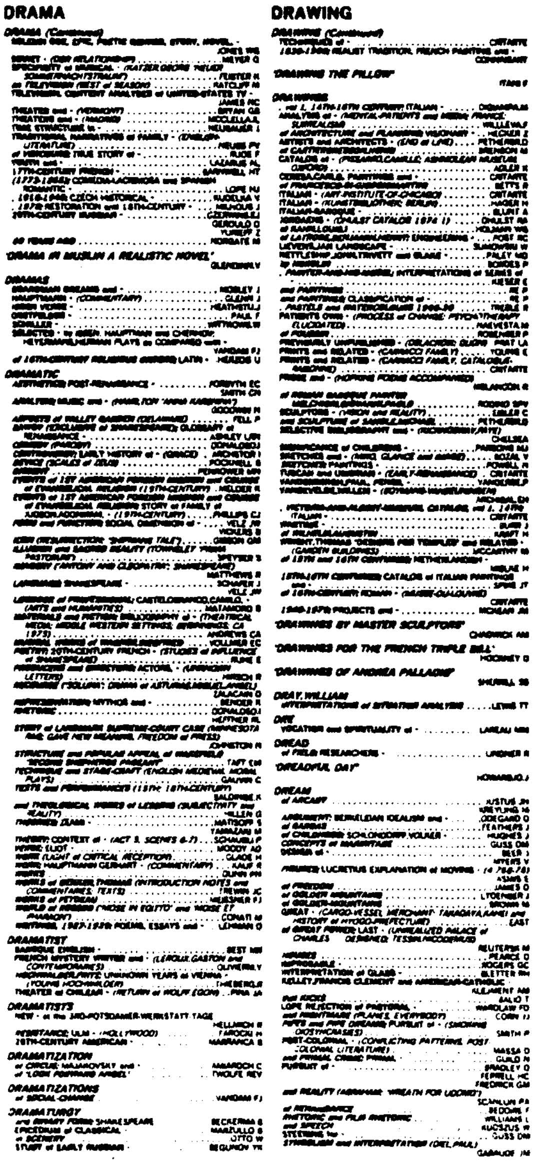

```
•INDEXING

•INDEXING (Continued)
  CLASSIFICATION and • (RELEVANCE of INTERCONCEPT)
                                              IC-80-2-579
  CONCEPTS and METHODS (BOOK REVIEW of BORKO,H.-
     BERNIER,CH.L.) ........................ IC-80-2-557
  CONVERGENCE and COMPATIBILITY ANALYSIS of ALL-
     UNION-BOOK-CHAMBER and MEDICAL LIBRARIES
     SUBJECT • ............................. IC-80-2-564
  CONVERGENCE; ANALYSIS of SUBJECT • (USE of
     CENTRALIZED SUBJECT INDEXING PERFORMED; LIBRARY-
     OF-THE-SOCIAL-SCIENCES-ACADEMY, CENTRAL-
     COMMITTEE-OF-THE-CPSU) ................ IC-80-2-583
  and CURRENT AWARENESS SERVICES; ARTICULATED
     SUBJECT • (COMPARISON; KWOC and PRECIS INDEXING,
     LIBRARY STUDIES) ...................... IC-80-2-581
  DATABASE • ............................... IC-80-2-580
  of DOCUMENTS; SUBJECT • (REALISATION of
     MATHEMATICAL MODELS) .................. IC-80-2-566
  of EARLY SCIENTIFIC PERIODICALS (INDEX CATALOGUE)
                                              IC-80-2-582
  EXPERIMENT; REPORT of • (BOOK INDEXING, TOOL,
     READING, BUTSUZO-KOKORO, KATACHI, TEXT)
                                              IC-80-2-581
  FAILURE; STUDY of • (THESAURUS-BASED AUTOMATIC
     INDEXING) ............................. IC-80-2-567
  HAWAIIAN-PACIFIC • ....................... IC-80-2-582
  KWOC and PRECIS • (ARTICULATED SUBJECT INDEXING and
     CURRENT AWARENESS SERVICES, COMPARISON, LIBRARY
     STUDIES) .............................. IC-80-2-581
  LANGUAGES PATTERNED; REDUNDANT • (NATURAL
     LANGUAGE) ............................. IC-80-2-560
  LANGUAGES; FOUNDATION of • (CLASSIFICATION PROCESS)
                                              IC-80-2-667
  LAW ENFORCEMENT and CRIMINAL JUSTICE INFORMATION
     (NATIONAL CRIMINAL JUSTICE THESAURUS,
     DESCRIPTORS) .......................... IC-80-2-550
  MACHINE-AIDED • (ENCOUNTER; AUTOMATIC INDEXING of
     3RD KIND) ............................. IC-80-2-567
  METHOD; REFERENCE STRING • ............... IC-80-2-581
  PAPER PRODUCTION and SYSTEM of • (APPLICATION of
     COMPUTERS, INTRODUCING, THESAURUS, BULLETIN)
                                              IC-80-2-561
  PERFORMED; USE of CENTRALIZED SUBJECT • (LIBRARY-OF-
     THE-SOCIAL-SCIENCES-ACADEMY, CENTRAL-COMMITTEE-
     OF-THE-CPSU; ANALYSIS of SUBJECT INDEXING
     CONVERGENCE) .......................... IC-80-2-583
  the PERIODICAL LITERATURE of ANGLO-AMERICAN
     BIBLIOGRAPHY .......................... IC-80-2-582
  of PRINTED DOCUMENTS; COMBINATION of CLASSIFICATION
     and SUBJECT • ......................... IC-80-2-565
  PROCESS; QUANTITATIVE DESCRIPTION of • ... IC-80-2-565
  and RESEARCH (APPLICATION of INFORMATION SCIENCE
     RECOGNITIONS, SPECIAL SCIENCES, INFORMATION
     SYSTEMS, LARGE-CAPACITY STORAGE) ...... IC-80-2-559
  and RETRIEVAL; INTELLIGENT • (MAN-MACHINE
     PARTNERSHIP) .......................... IC-80-2-565
  ROLE of AUTOMATIC • (OPERATIONAL ONLINE RETRIEVAL
     SYSTEMS) .............................. IC-80-2-554
  STRING • (MICROCOMPUTER-GENERATED GRAPHIC
     DISPLAYS, AID) ........................ IC-80-2-580
  SURVEY of AUSTRALIAN • (OBSERVATIONS) .... IC-80-2-584
  SYSTEM; COMPUTER-ASSISTED DYNAMIC • (CADIS)
                                              IC-80-2-567
  SYSTEM; THESAURUS and • (PROBLEMS of UNIVERSITY
     ORGANIZATION) ......................... IC-80-2-575
  TERMINOLOGY POSTING TERMS; DOC RETRIEVAL and •
     (HIERARCHY and KWOC) .................. IC-80-2-548
  and THEORETICAL FOUNDATIONS; PRACTICE of •
                                              IC-80-2-565
  THESAURUS-BASED AUTOMATIC • (STUDY of INDEXING
     FAILURE) .............................. IC-80-2-567
  THESAURUS-BASED DOCUMENT • (RECOGNITION of
     MULTICOMPONENT TERMS) ................. IC-80-2-576
  by UDC (PRODUCTION QUALITY FIELD) ........ IC-80-2-568
  UNIFICATION; PROBLEMS of COORDINATE • (NETWORK of
     AUTOMATED STI CENTRES) ................ IC-80-2-584
  VERSUS KWOC; AUTOMATED • (PERFORMANCE
     COMPARISON) ........................... IC-80-2 568
  VOCABULARY; FREE • ....................... IC-80-2 566
  from 1957-2000; TRENDS in • .............. IC-80-2-553
  of 3RD KIND; AUTOMATIC • (MACHINE-AIDED INDEXING
     ENCOUNTER) ............................ IC-80-2 567

•INDUSTRIAL
  ENTERPRISES (USE of PATENT INFORMATION and
     INTERNATIONAL PATENT CLASSIFICATION) .. IC-80-2-572
  INFORMATION THESAURUS (CENTRAL-AMERICA and
     DOMINICAN-REPUBLIC) ................... IC-80-2 550
```

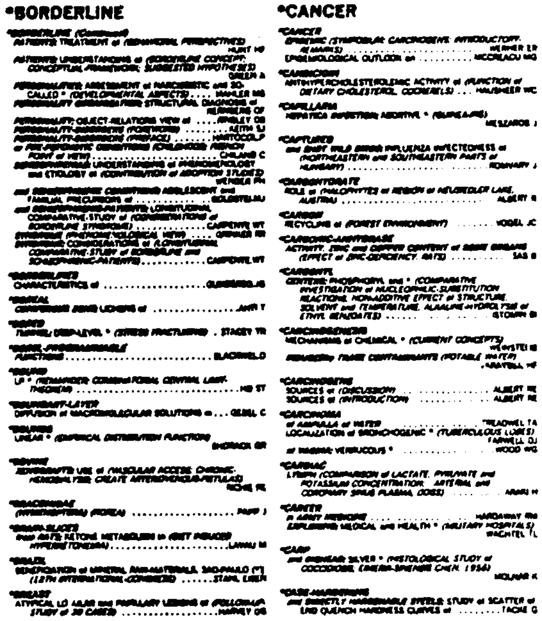
Figure 8
A photocomposed KWPSI sample showing details of its structure and appearance.

Index" includes besides titles of articles, also titles of book reviews, as well as descriptions of musical performances and musical records, compiled according to special rules. Many contain multiword names of works of art, taken in quotation marks, which have to be handled as single words. A KWPSI sample for arts and humanities is shown on figure 9. Two more KWPSI samples are given on figure 10 (science and technology titles) and figure 11 (geoscience research front names).


DRAMA

DRAWING

Figure 9
KWPSI samples for arts and humanities titles.


•BORDERLINE

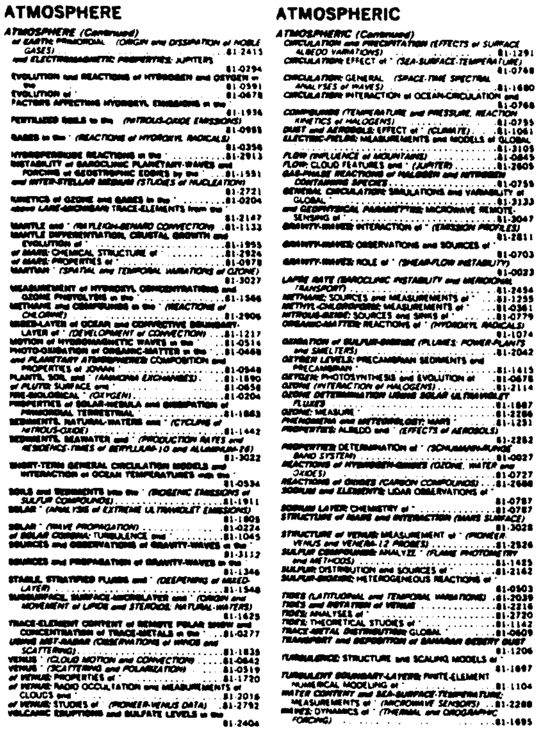
•CANCER

Figure 10
KWPSI sample for titles covered by the multidisciplinary SCI (science and technology).

3. Possible usage of automatically detected word phrases for enhanced online access to databases

The common method of online access to commercially available textual databases of both bibliographic and full text type is through boolean queries formulated in terms of single words. Automatically detectable word phrases of the type used in the KWPSI system could be used in three different ways for improving online access.

One extreme way would involve the creation of word phrases of the above type at the input stage for every informative word of the input. In response to a single word query a sequence of screens would be shown displaying the image of what in a printed KWPSI would be the KWPSI section under the given word taken as heading. After browsing online some part of this up-to-date online SI the user could choose to limit further browsing by responding with an additional search term, most likely chosen from some of the already examined index entries. As a result the system would reply by eliminating from the displayed output the entries not containing the given word and the user would continue to browse the so trimmed display. Several such iterations could be performed



ATMOSPHERE

ATMOSPHERE (Continued)
of EARTH: PRIMORDIAL (ORIGIN and DISSIPATION of NOBLE GASES) 81-2415
and ELECTROMAGNETIC PROPERTIES: JUPITER'S 81-0294
EVOLUTION and REACTIONS of HYDROGEN and OXYGEN in the 81-0591
EVOLUTION of ' 81-0678
FACTORS AFFECTING HYDROXYL EMISSIONS in the ' 81-1936
FERTILIZED SOILS to the (NITROUS-OXIDE EMISSIONS) 81-0988
GASES in the ' (REACTIONS of HYDROXYL RADICALS) 81-0358
HYDROPEROXIDE REACTIONS in the 81-2913
INSTABILITY of BAROCLINIC PLANETARY-WAVES and FORCING of GEOSTROPHIC EDDIES by the 81-1551
and INTER-STELLAR MEDIUM (STUDIES of NUCLEATION) 81-2721
KINETICS of OZONE and GASES in the 81-0204
above LAKE-MICHIGAN: TRACE-ELEMENTS from the 81-2147
MANTLE and ' (RAYLEIGH-BENARD CONVECTION) 81-1133
MANTLE DIFFERENTIATION, CRUSTAL GROWTH and EVOLUTION of ' 81-1995
of MARS: CHEMICAL STRUCTURE of 81-2924
of MARS: PROPERTIES of ' 81-0978
MARTIAN ' (SPATIAL and TEMPORAL VARIATIONS of OZONE) 81-3027
MEASUREMENT of HYDROXYL CONCENTRATIONS and OZONE PHOTOLYSIS in the 81-1546
METHANE and COMPOUNDS in the ' (REACTIONS of CHLORINE) 81-2906
MIXED-LAYER of OCEAN and CONVECTIVE BOUNDARY-LAYER of ' (DEVELOPMENT of CONVECTION) 81-1217
MOTION of HYDROMAGNETIC WAVES in the 81-0514
PHOTO-OXIDATION of ORGANIC-MATTER in the 81-0468
and PLANETARY ATMOSPHERES: COMPOSITION and PROPERTIES of JOVIAN ' 81-0948
PLANTS, SOIL and ' (AMMONIA EXCHANGES) 81-1890
of PLUTO: SURFACE and ' 81-0658
PRE-BIOLOGICAL ' (OXYGEN) 81-0204
PROPERTIES of SOLAR-NEBULA and DISSIPATION of PRIMORDIAL TERRESTRIAL 81-1883
SEDIMENTS, NATURAL-WATERS and ' (CYCLING of NITROUS-OXIDE) 81-1442
SEDIMENTS, SEAWATER and ' (PRODUCTION RATES and RESIDENCE-TIMES of BERYLLIUM-10 and ALUMINUM-26) 81-3022
SHORT-TERM GENERAL CIRCULATION MODELS and INTERACTION of OCEAN TEMPERATURES with the ' 81-0534
SOILS and SEDIMENTS into the ' (BIOGENIC EMISSIONS of SULFUR COMPOUNDS) 81-1911
SOLAR ' (ANALYSIS of EXTREME ULTRAVIOLET EMISSIONS) 81-1805
SOLAR ' (WAVE PROPAGATION) 81-0274
of SOLAR CORONA: TURBULENCE and ' 81-1045
SOURCES and OBSERVATIONS of GRAVITY-WAVES in the ' 81-3112
SOURCES and PROPAGATION of GRAVITY-WAVES in the 81-1346
STABLE, STRATIFIED FLUIDS and ' (DEEPENING of MIXED-LAYER) 81-1548
SUBSURFACE, SURFACE-MICROLAYER and ' (ORIGIN and MOVEMENT of LIPIDS and STEROIDS: NATURAL-WATERS) 81-1625
TRACE-ELEMENT CONTENT of REMOTE POLAR SNOW and CONCENTRATION of TRACE-METALS in the ' 81-0277
USING MST-RADAR (OBSERVATIONS of WINDS and SCATTERING) 81-1835
VENUS ' (CLOUD MOTION and CONVECTION) 81-0642
VENUS ' (SCATTERING and POLARIZATION) 81-0519
of VENUS: PROPERTIES of ' 81-1720
of VENUS: RADIO OCCULTATION and MEASUREMENTS of CLOUDS and ' 81-2016
of VENUS: STUDIES of ' (PIONEER-VENUS DATA) 81-2792
VOLCANIC ERUPTIONS and SULFATE LEVELS in the 81-2404

ATMOSPHERIC

ATMOSPHERIC (Continued)
CIRCULATION and PRECIPITATION (EFFECTS of SURFACE ALBEDO VARIATIONS) 81-1291
CIRCULATION: EFFECT of ' (SEA-SURFACE-TEMPERATURE) 81-0768
CIRCULATION: GENERAL (SPACE-TIME SPECTRAL ANALYSES of WAVES) 81-1680
CIRCULATION: INTERACTION of OCEAN-CIRCULATION and 81-0768
COMPOUNDS (TEMPERATURE and PRESSURE, REACTION KINETICS of HALOGENS) 81-0795
DUST and AEROSOLS: EFFECT of ' (CLIMATE) 81-1061
ELECTRIC-FIELDS: MEASUREMENTS and MODELS of GLOBAL 81-3105
FLOW (INFLUENCE of MOUNTAINS) 81-0845
FLOW: CLOUD FEATURES and ' (JUPITER) 81-2605
GAS-PHASE REACTIONS of HALOGEN and NITROGEN CONTAINING SPECIES 81-0755
GENERAL CIRCULATION: SIMULATIONS and VARIABILITY of GLOBAL ' 81-3133
and GEOPHYSICAL PARAMETERS: MICROWAVE REMOTE-SENSING of ' 81-3047
GRAVITY-WAVES: INTERACTION of ' (EMISSION PROFILES) 81-2811
GRAVITY-WAVES: OBSERVATIONS and SOURCES of ' 81-0703
GRAVITY-WAVES: ROLE of ' (SHEAR-FLOW INSTABILITY) 81-0023
LAPSE RATE (BAROCLINIC INSTABILITY and MERIDIONAL TRANSPORT) 81-2494
METHANE: SOURCES and MEASUREMENTS of ' 81-1255
METHYL-CHLOROFORM: MEASUREMENTS of ' 81-0361
NITROUS-OXIDE: SOURCES and SINKS of ' 81-0779
ORGANIC-MATTER: REACTIONS of ' (HYDROXYL RADICALS) 81-1074
OXIDATION of SULFUR-DIOXIDE (PLUMES: POWER-PLANTS and SMELTERS) 81-2042
OXYGEN LEVELS: PRECAMBRIAN SEDIMENTS and PRECAMBRIAN ' 81-1415
OXYGEN: PHOTOSYNTHESIS and EVOLUTION of 81-0678
OZONE (INTERACTION of HALOGENS) 81-2114
OZONE DETERMINATION USING SOLAR ULTRAVIOLET FLUXES 81-1887
OZONE: MEASURE ' 81-2286
PHENOMENA and METEOROLOGY: MARS ' 81-1231
PROPERTIES: ALBEDO and ' (EFFECTS of AEROSOLS) 81-2282
PROPERTIES: DETERMINATION of ' (SCHUMANN-RUNGE BAND SYSTEM) 81-0027
REACTIONS of HYDROGEN-OXIDES (OZONE, WATER and OXIDES) 81-0727
REACTIONS of OXIDES (CARBON COMPOUNDS) 81-2686
SODIUM and ELEMENTS: LIDAR OBSERVATIONS of ' 81-0787
SODIUM LAYER: CHEMISTRY of ' 81-0787
STRUCTURE of MARS and INTERACTION (MARS SURFACE) 81-3028
STRUCTURE of VENUS: MEASUREMENT of ' (PIONEER-VENUS and VENERA-12 PROBES) 81-2326
SULFUR COMPOUNDS: ANALYZE ' (FLAME PHOTOMETRY and METHODS) 81-1425
SULFUR: DISTRIBUTION and SOURCES of ' 81-2142
SULFUR-DIOXIDE: HETEROGENEOUS REACTIONS of ' 81-0503
TIDES (LATITUDINAL and TEMPORAL VARIATIONS) 81-2039
TIDES and ROTATION of VENUS 81-2216
TIDES: ANALYSES of ' 81-2720
TIDES: THEORETICAL STUDIES of ' 81-1142
TRACE-METAL DISTRIBUTION: GLOBAL ' 81-0609
TRANSPORT and DEPOSITION of SAHARAN DESERT DUST 81-1206
TURBULENCE: STRUCTURE and SCALING MODELS of ' 81-1097
TURBULENT BOUNDARY-LAYERS: FINITE-ELEMENT NUMERICAL MODELING of ' 81-1104
WATER CONTENT and SEA-SURFACE-TEMPERATURE: MEASUREMENTS of ' (MICROWAVE SENSORS) 81-2288
WAVES: DYNAMICS of ' (THERMAL and OROGRAPHIC FORCING) 81-1095

Figure 11
KWPSI sample for names of geoscience research fronts.

until the user would be left with the display of a SI to relevant items of the database. This SI would be then printed together with the full list of relevant items. It is thought that such kind of interaction could be more user friendly than the currently used boolean mode.

Another way of using the KWPSI technique in an online environment would be to use the KWPSI format for the output of the results of a retrieval performed in a traditional boolean way. The query word which achieved the most strong trimming effect would be used as heading.

A third way would involve the compression of a KWPSI section under a given heading before it is displayed in response to a word. One could e.g. retain only such noun phrases containing the given word which occur at least k times in the database. An example of such list for the "Arts and Humanities" database is given in figure 12. By displaying such lists of words closely co- occurring with a given one the system would perform thesaurus-type functions.

The implementation of all such possibilities would be rather difficult for any existing system in view of the effort required to reprocess past input. Instead, after the input of a query word the corresponding full text records could be called a not processed online in core for generating KWPSI-type index records. In this case all the above functions could be still performed.

*<u>FILM(S)</u>

AFRICAN *
AMERICAN *
ARCHIVES
AVANT-GARDE *
BLACK *
BRAZILIAN *
COMPANY
CRITICISM
DIRECTOR
DOCUMENTARY *
FESTIVAL(S)
HISTORY
HORROR(S) *
INDUSTRY
LITERARY *
MAKER(S)
OBSCENE *
POLISH *
PRODUCTION
PROPAGANDA
TECHNIQUES
THEORY
TV *
WESTERN *

Figure 12
List of word phrases containing the word 'FILM(S)' and appearing at least twice in titles covered during a three months period by ISI's ARTS and HUMANITIES CITATION INDEX. Such automatically compiled lists are suggested as search aides for the online access to databases.

Another possibility which we are considering is to place the KWPSI-type processing capabilities into a microcomputer which is being used to mediate online searches in remote databases. All the text records containing a given (not too frequent) word could be initially tapped from the database into the microcomputer. Following that the microcomputer could perform all the above functions in an offline mode.

REFERENCES

Armitage, J.E., Lynch, M.F. "Articulation in the Generation of Subject Indexes by Computer." <u>Journal of Chemical Documentation</u>: 7:170-8, 1967.

Cohen, S.M., Dayton, D.L., Salvador, R. "Experimental Algorithmic Generation of Articulated Index Entries from Natural Language Phrases at Chemical Abstracts Service." <u>Journal of Chemical Information and Computer Sciences</u>: 1976 May, 16(2): 93-99.

Garfield, E. "The Permuterm Subject Index: An Autobiographic Review." <u>Journal of the American Society for Information Science</u>: 27(5/6): 288-291, 1976.

Klein S., Simmons R.F. "A Computational Approach to grammatical coding of English words." <u>Journal of ACM</u>: 10(3):334-347, 1963.

Luhn P. "Keyword-in-Context Index for Technical Literature." Report RC 127. New York: IBM Corp., Advanced System Development Division, 1959.

Vladutz G., Garfield E. "KWPSI - An Algorithmically derived Key Word/Phrase Subject Index." <u>Proceedings of the ASIS, 42nd Annual Meeting</u>, Minneapolis, Minnesota, October 14-18, 1979; pp. 236-245.