ANNOD -- A Navigator of Natural-language Organized (Textual) Data

Robert E. Williamson
Knowledge Systems, Inc.
12 E. Melrose St.
Chevy Chase, MD 20815

Work performed at:
Information Technology Branch, LHNCBC
National Library of Medicine
Bethesda, MD 20209

ABSTRACT

ANNOD is the name of a system developed at the National Library of Medicine (NLM), which implements a set of linguistic and empirical techniques that permit retrieval of natural language information in response to natural language queries. The system is based on Dr. Gerard Salton's SMART [1] document retrieval system and is presently implemented on a mini-computer as part of an Interactive TExt Management System, ITEMS.[2] Actual experience with retrieval of information from NLM's Hepatitis Knowledge Base (HKB), an encyclopedic hierarchical, full-text file, is presented. The techniques used in ANNOD include: automatic stemming of words, common word deletion, thesaurus expansion, a complex empirical matching (ranking) algorithm (similarity measure), and techniques expressly designed to permit rapid response in a mini-computer environment. Preliminary testing demonstrates high efficiency in identifying portions of a text which are relevant to users.

I. INTRODUCTION

Full text data and knowledge bases are becoming increasingly available in machine-readable form. Computers are being used to process full text and to retrieve needed information. Methods which enable untrained users to economically locate unindexed information are needed. The most expensive aspects of most retrieval systems are their dependence on highly trained indexers and search specialists. Enabling a user to enter unrestricted natural language queries would make retrieval systems easier to use. Such queries should elicit responses and present text items for perusal which contain information relevant to the query while requiring little or no further decision making on the part of the user.

"A Navigator of Natural Language Organized Data" (ANNOD) is a recently develped retrieval system which combines the use of linguistically and empirically derived parameters to rank individual paragraphs of full text for their similarity to natural language queries submitted by users. The system is based in Dr. Gerard Salton's SMART [1] document retrieval system and is presently implemented on a mini-computer as part of an "Interactive TExt Management System", ITEMS. [2]

ANNOD was initially developed for the Hepatitis Knowledge Base (HKB), A Prototype Information Transfer System, [3] a hierarchical full text file that was available at the Lister Hill National Center for Biomedical Communications. The central features of ANNOD are generic and are expected to have considerable universality. ANNOD should be readily transferable to textual material in other content domains such as law, business, history, literature, etc. The core of ANNOD is an empirical algorithm; specific features have been added that reflect details of the HKB's organization and structure.

[1] Salton, G. The SMART Retrieval System, Prentice-Hall, 1971.

[2] Williamson, Robert. Experience With and Plans for Extending an Interactive Text Management System, ITB, National Library of Medicine, Bethesda, MD., to be published.

[3] Bernstein, L. M., Siegel, E. R., Goldstein, C. M. "The Hepatitis Knowledge Base -- A Prototype Information Transfer System", ANNALS of Internal Medicine, 93: July 1980, Part 2, 169-181.

The purpose of this paper is to present ANNOD, a technique for end-user access to textual information, in sufficient detail to permit implementation of a comparable system. To do so requires a description of its operation in the Hepatitis Knowledge Base environment, the only full-text file to which it has been applied. Some of the empirical factors used for the HKB will be altered significantly in subsequent applications of ANNOD to other texts. By changing internal parameters and by adding new features based on the unique structure of the HKB, successive versions of the algorithm have shown progressive improvement. Improvement resulted from crucial interaction and complementarity of the system designer's (Robert E. Williamson) computer science expertise with that of Dr. Lionel M. Bernstein, the physician who developed the initial Hepatitis Knowledge Base. (The degree of success of the current version warrants its reporting at this time; however, progressive improvements in later versions are anticipates.)

Section II of this paper provides a brief overview of the environment in which ANNOD has been installed. The next section describes ANNOD as the end-user encounters it. Sections IV and V describe the algorithm itself in sufficient detail to permit implementation of a comparable system. The remaining sections discuss modifications already made and those planned in the near future. Portions of the full 46 page paper (as noted by * in the Contents below) are cut significantly or eliminated entirely. Please write the author to obtain a complete copy of this paper, [3] or [5].

CONTENTS

II. THE ENVIRONMENT IN WHICH ANNOD OPERATES

A. The Hepatitis Knowledge Base and ITEMS

While the ideas encompassed by the label ANNOD provide a generic natural language query capability, their initial implementation has been to provide access for the Hepatitis Knowledge Base (HKB). [3] The HKB is similar to an encyclopedia or text book -- namely paragraphs of textual information at varying levels of detail -- and contains over 2000 paragraphs (more than one million characters of text) plus references and tables. The Interactive TExt Management System (ITEMS) is the software system developed to create, maintain and provide (limited) access to the HKB.

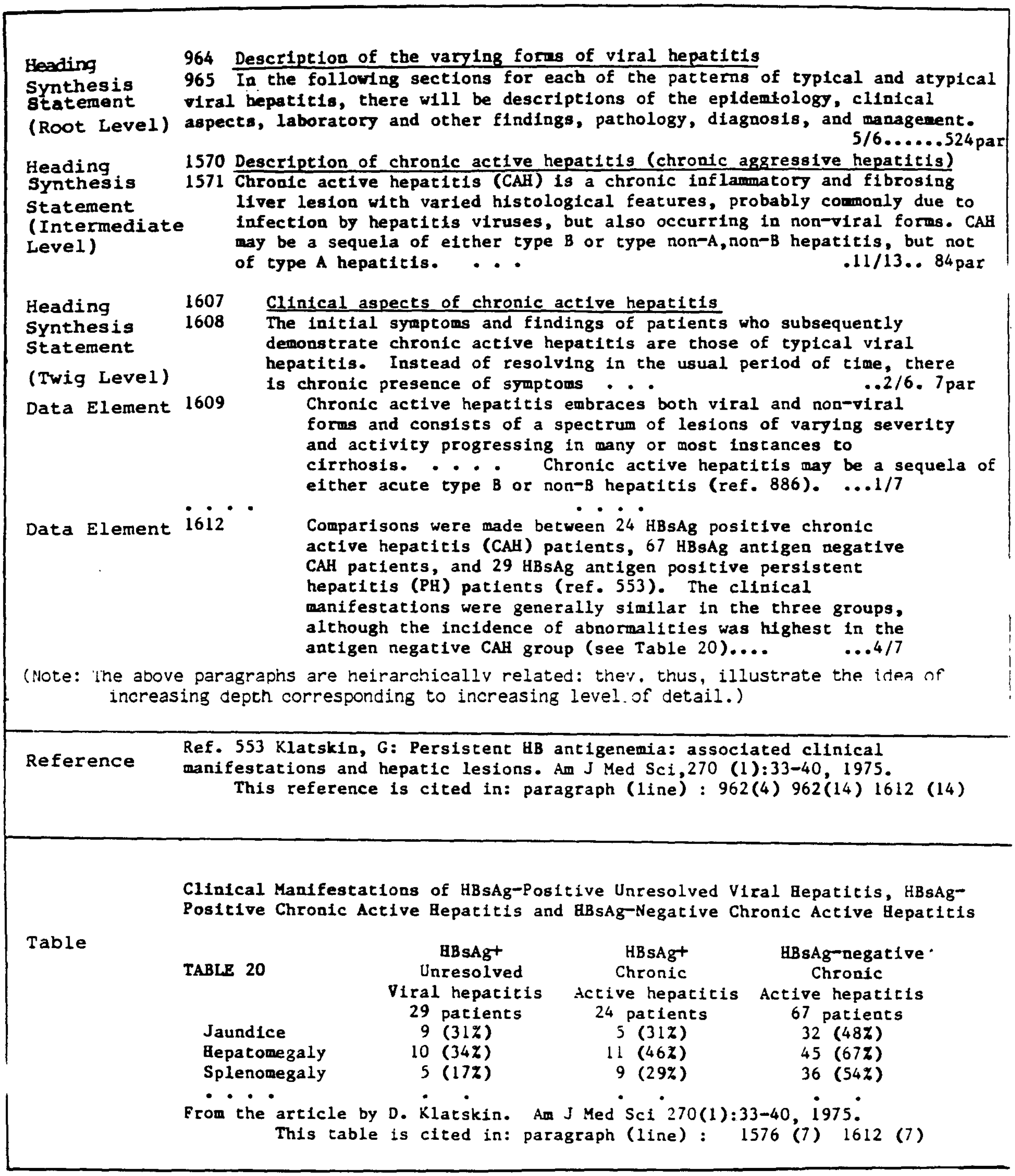
The Hepatitis Knowledge Base (HKB) has been extensively described [3] and studied in limited field trials [4]. The description of the HKB here is limited to that necessary for understanding the methods used by ANNOD. The HKB was the result of a multi-disciplinary team effort to assemble a compact yet comprehensive body of information that could meet health practitioners' needs. The aim was to assemble knowledge on viral hepatitis which would contain substantive information relevant to a wide variety of questions (rather than bibliographic citations only); would provide information that is both current and the consensus of a group of experts; would be immediately responsive to inquiries (on-line access); and would provide access to variable levels of data supporting the substantive information, including citations to primary publications for more detailed study when desired.

1. A Top-Down View. The information in the HKB, see examples in Figure 1, is arranged in a hierarchy of topic headings and series of subheadings. The topic headings comprise a detailed listing of the material one would expect in a large, thorough textbook on hepatitis and are functionally used as a Table of Contents. For each topic heading or subheading there is a succeeding paragraph at the same margin that is a "synthesis" of the state of knowledge about that heading. In turn, each synthesis paragraph is supported by further indented, highly selective "data paragraphs" derived from experts' previously published (quality filtered) source documents. Citations included in the data paragraphs are to the primary publications referenced by the experts in their published articles to support their statements.

2. A Bottom-Up View. The most detailed level of the HKB consists of paraphrases of published material which are called "detail elements paragraphs" (Figure 1). The major findings contained within each group of content related detail elements are summarized by two "paragraphs" that always occur as a pair: a heading and a synthesis-statement -- paragraphs at this level are known as "twigs". A series of related heading/synthesis-statement pairs are frequently grouped under a broader heading/synthesis-statement pair -- such a pair is one level higher in the data base.

[4] Roderer, NK, King, DW, McDonald, DD, and Bush, CG. Evaluation of the Hepatitis Knowledge Base. Rockville, MD: King Research, Inc.: 1981 (Lister Hill National Center for Biomedical Communications, National Library of Medicine) NTIS Order #82-199639.

Rev. 9/7/83



Heading 964 Description of the varying forms of viral hepatitis

Synthesis Statement (Root Level) 965 In the following sections for each of the patterns of typical and atypical viral hepatitis, there will be descriptions of the epidemiology, clinical aspects, laboratory and other findings, pathology, diagnosis, and management. 5/6......524par

Heading 1570 Description of chronic active hepatitis (chronic aggressive hepatitis)

Synthesis Statement (Intermediate Level) 1571 Chronic active hepatitis (CAH) is a chronic inflammatory and fibrosing liver lesion with varied histological features, probably commonly due to infection by hepatitis viruses, but also occurring in non-viral forms. CAH may be a sequela of either type B or type non-A,non-B hepatitis, but not of type A hepatitis. . . . .11/13.. 84par

Heading 1607 Clinical aspects of chronic active hepatitis

Synthesis Statement (Twig Level) 1608 The initial symptoms and findings of patients who subsequently demonstrate chronic active hepatitis are those of typical viral hepatitis. Instead of resolving in the usual period of time, there is chronic presence of symptoms 2/6. 7par

Data Element 1609 Chronic active hepatitis embraces both viral and non-viral forms and consists of a spectrum of lesions of varying severity and activity progressing in many or most instances to cirrhosis. . . . . Chronic active hepatitis may be a sequela of either acute type B or non-B hepatitis (ref. 886). ...1/7

. . . .

Data Element 1612 Comparisons were made between 24 HBsAg positive chronic active hepatitis (CAH) patients, 67 HBsAg antigen negative CAH patients, and 29 HBsAg antigen positive persistent hepatitis (PH) patients (ref. 553). The clinical manifestations were generally similar in the three groups, although the incidence of abnormalities was highest in the antigen negative CAH group (see Table 20).... ...4/7

(Note: The above paragraphs are heirarchicallv related: thev, thus, illustrate the idea of increasing depth corresponding to increasing level of detail.)

Reference Ref. 553 Klatskin, G: Persistent HB antigenemia: associated clinical manifestations and hepatic lesions. Am J Med Sci,270 (1):33-40, 1975.
This reference is cited in: paragraph (line) : 962(4) 962(14) 1612 (14)

Table Clinical Manifestations of HBsAg-Positive Unresolved Viral Hepatitis, HBsAg-Positive Chronic Active Hepatitis and HBsAg-Negative Chronic Active Hepatitis

TABLE 20	HBsAg+ Unresolved Viral hepatitis 29 patients	HBsAg+ Chronic Active hepatitis 24 patients	HBsAg-negative Chronic Active hepatitis 67 patients
Jaundice	9 (31%)	5 (31%)	32 (48%)
Hepatomegaly	10 (34%)	11 (46%)	45 (67%)
Splenomegaly	5 (17%)	9 (29%)	36 (54%)
. . . .	. .	. .	. .

From the article by D. Klatskin. Am J Med Sci 270(1):33-40, 1975.
This table is cited in: paragraph (line) : 1576 (7) 1612 (7)

Figure 1 — Portions of the Hepatitis Knowledge Base

Currently, the hierarchy has a maximum of six levels with an average of three levels (above the detail elements). The "....#1/#2... #3par" sequences after paragraphs in the figures are an HKB feature that shows how deep in the tree the paragraph is (the number of leading periods), the index of the paragraph (#1) relative to other siblings at the same level (#2), the maximum depth of paragraphs (number of hierarchical levels) below the given unit (the number of trailing periods), and the number of paragraphs below the displayed unit (#3).

B. The MUMPS Programming Environment

ANNOD and ITEMS are implemented in the MIIS dialect of the MUMPS programming language on a relatively small Data General Eclipse minicomputer (a C-330 with 128 K bytes of memory), and are accessible via a dial-up line or via Telenet from any ASCII terminal. MUMPS is a powerful, high-level interpreted, language which supports easy development of interactive systems. MUMPS directly supports a single-level data store called a "global"; in fact, one cannot directly request any disk I/O. In addition, the MUMPS environment includes significant text processing primitives and support for arrays with any number of variable-length alphabetic subscripts -- data fields can also consist of multiple, variable-length alphabetic fields.

1. **Should physicians with HbsAg positive tests treat patients?**
2. **Damage to cardiopulmonary system?**
3. **What is the ecology of HBV?**
4. **Pathological criteria for differentiating Type A from B?**
5. **What subtypes are found in Japan?**
6. HBV risk to teachers in schools for the mentally handicapped?
7. How many people contracted Type A hepatitis in 1980?
8. Are dialysis patients with chronic kidney disease affected?
9. What are the pathological findings of viral hepatitis on the acinus?
10. Ultrastructural morphology of HBV?
11. Transmission modes in homosexuals?
12. What is the relationship between stool shedding of hepatitis A Virus and symptoms of Type A hepatitis?
13. Can mothers infect their babies
14. Localization of antigens in cells during necrosis
15. Incidence of cirrhosis in hcc hepatocellular carcinoma
16. Role of bed rest or exercise in management
17. treatment of type A hepatitis with gamma globulin
18. Is ISG useful in treating hepatitis B infection

Figure 2 -- Actual Hepatitis Knowledge Base Queries Submitted to ANNOD

III. THE USER'S VIEW

ANNOD provides a user with a cordial method of accessing information. The user usually is not concerned with how ANNOD works (Sections IV and V); he or she is typically interested in only two features of ANNOD: 1) ANNOD accepts any natural language expression, and 2) the extent to which ANNOD enables the user to satisfy his/her information needs with less effort and/or greater accuracy than with other techniques.

A user's query can consist of any English language description; typical queries are shown in Figure 2. ANNOD then locates the "most highly related" paragraphs (per criteria presented in Sections IV and V). The user is provided with a highlighted display of the "best" paragraph (Figure 3c) and a summary indication of a few highly related paragraphs, (Figure 4). The user can obtain extensions to the summary display. The researcher or retrieval specialist can obtain a table which explicitly indicates the links between each specified paragraph and the query (Figure 5). (A detailed explanation of the content of Figures 3-5 is given in the sections below.)

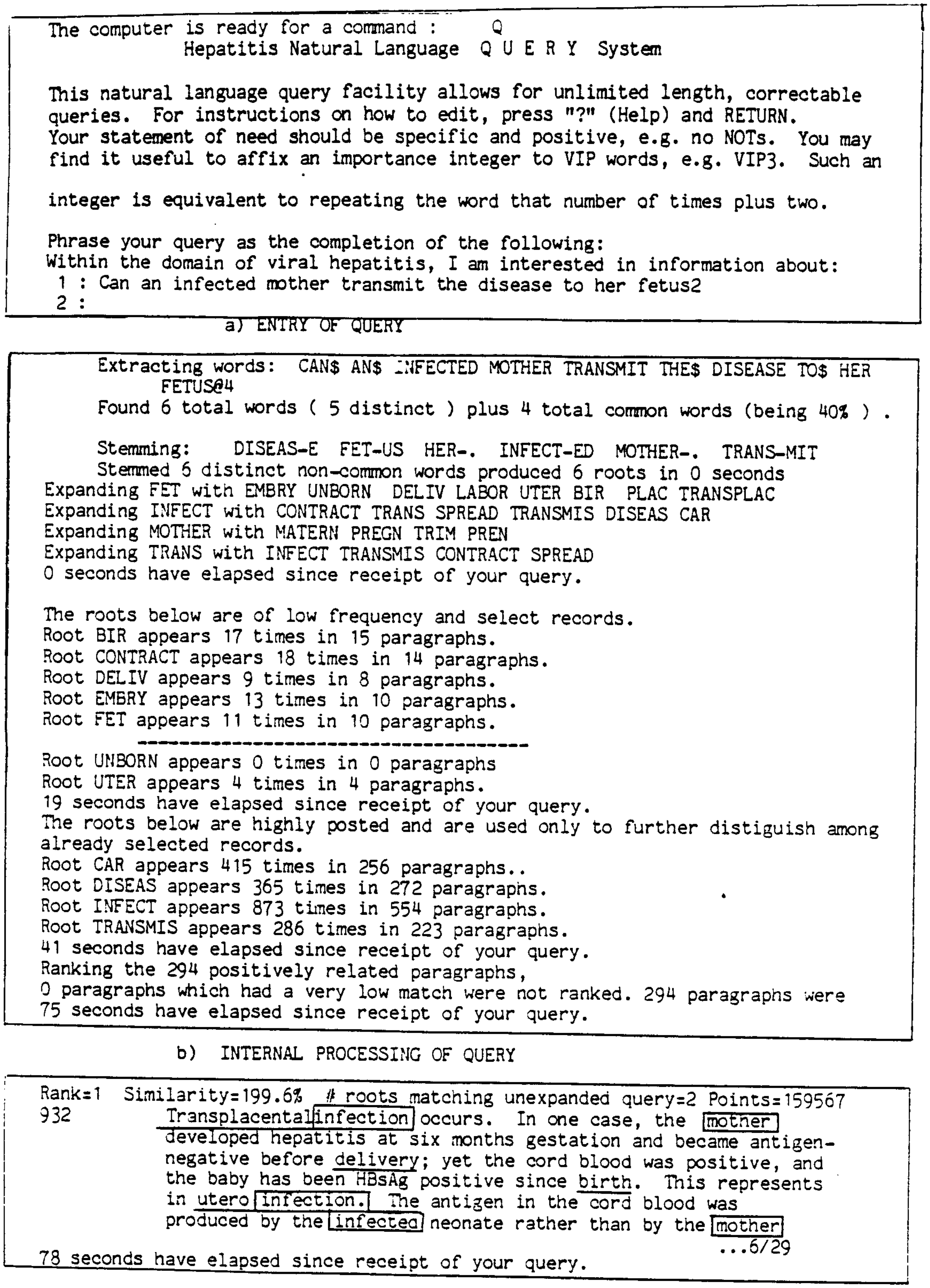
A user is returned to ITEMS after each use of ANNOD and thus retains all of the browsing flexibility provided by the ITEMS software. Interactions with ITEMS are unrestricted (i.e., any command in Table A-1 may be used) and include: 1) display of contextually related paragraphs, e.g., parents and children, and 2) display of full citations of the articles cited to support a given statement. Subsequent use of ANNOD usually consists of requests for the "next best undisplayed" paragraph, requests for information based on a revised query, or totally new queries.

A. The Highlighted ITEMS Display

ANNOD presents each retrieved paragraph with the words which led to retrieval highlighted (relative to other words in the paragraph). Highlighting on a CRT, see Figure 3c, typically consists of inverse video for words present in the user's entered query (indicated by boxed words in this example). Words related to the user's query via the system-provided thesaurus are underlined. On terminals which do not support such highlighting, flag characters precede each normally highlighted word, presently ">>" for words in a user's query and "__" for thesaurus related words.

Figure 3 shows the detailed output associated with the processing of one query. In Figure 3a, the "Q" is the ITEMS command that invokes ANNOD. The query is then entered with the word "fetus" receiving extra weight (via the "2"). Figure 3b shows the research-oriented step-by-step processing of ANNOD; this output would be much briefer for typical users. In the line labeled "Extracting Words:" a "$" after a word denotes a common word that is ignored by ANNOD. The line labeled "Stemming" shows the stem of each word. The line labeled "Expanding" shows the roots added to the query based on substitution patterns in the thesaurus. (The HKB thesaurus was created by a physician in three days and contains only about 100 patterns. It is nonetheless estimated to provide 90% of the expansions useful for searching the HKB.)

Rev. 9/7/83



```
The computer is ready for a command :      Q
             Hepatitis Natural Language  Q U E R Y  System

This natural language query facility allows for unlimited length, correctable
queries.  For instructions on how to edit, press "?" (Help) and RETURN.
Your statement of need should be specific and positive, e.g. no NOTs.  You may
find it useful to affix an importance integer to VIP words, e.g. VIP3.  Such an

integer is equivalent to repeating the word that number of times plus two.

Phrase your query as the completion of the following:
Within the domain of viral hepatitis, I am interested in information about:
 1 : Can an infected mother transmit the disease to her fetus2
 2 :
```

a) ENTRY OF QUERY

```
     Extracting words:  CAN$ AN$ INFECTED MOTHER TRANSMIT THE$ DISEASE TO$ HER
           FETUS@4
     Found 6 total words ( 5 distinct ) plus 4 total common words (being 40% ) .

     Stemming:     DISEAS-E  FET-US  HER-.  INFECT-ED  MOTHER-.  TRANS-MIT
     Stemmed 6 distinct non-common words produced 6 roots in 0 seconds
Expanding FET with EMBRY UNBORN  DELIV LABOR UTER BIR  PLAC TRANSPLAC
Expanding INFECT with CONTRACT TRANS SPREAD TRANSMIS DISEAS CAR
Expanding MOTHER with MATERN PREGN TRIM PREN
Expanding TRANS with INFECT TRANSMIS CONTRACT SPREAD
0 seconds have elapsed since receipt of your query.

The roots below are of low frequency and select records.
Root BIR appears 17 times in 15 paragraphs.
Root CONTRACT appears 18 times in 14 paragraphs.
Root DELIV appears 9 times in 8 paragraphs.
Root EMBRY appears 13 times in 10 paragraphs.
Root FET appears 11 times in 10 paragraphs.
          -------------------------------------
Root UNBORN appears 0 times in 0 paragraphs
Root UTER appears 4 times in 4 paragraphs.
19 seconds have elapsed since receipt of your query.
The roots below are highly posted and are used only to further distiguish among
already selected records.
Root CAR appears 415 times in 256 paragraphs..
Root DISEAS appears 365 times in 272 paragraphs.
Root INFECT appears 873 times in 554 paragraphs.
Root TRANSMIS appears 286 times in 223 paragraphs.
41 seconds have elapsed since receipt of your query.
Ranking the 294 positively related paragraphs,
0 paragraphs which had a very low match were not ranked. 294 paragraphs were
75 seconds have elapsed since receipt of your query.
```

b) INTERNAL PROCESSING OF QUERY

```
Rank=1  Similarity=199.6%  # roots matching unexpanded query=2 Points=159567
932          Transplacental infection occurs.  In one case, the mother
             developed hepatitis at six months gestation and became antigen-
             negative before delivery; yet the cord blood was positive, and
             the baby has been HBsAg positive since birth.  This represents
             in utero infection.  The antigen in the cord blood was
             produced by the infected neonate rather than by the mother
                                                                 ...6/29
78 seconds have elapsed since receipt of your query.
```

c) DISPLAY OF A PRIORI BEST PARAGRAPH

Figure 3. Sample Natural Language Search

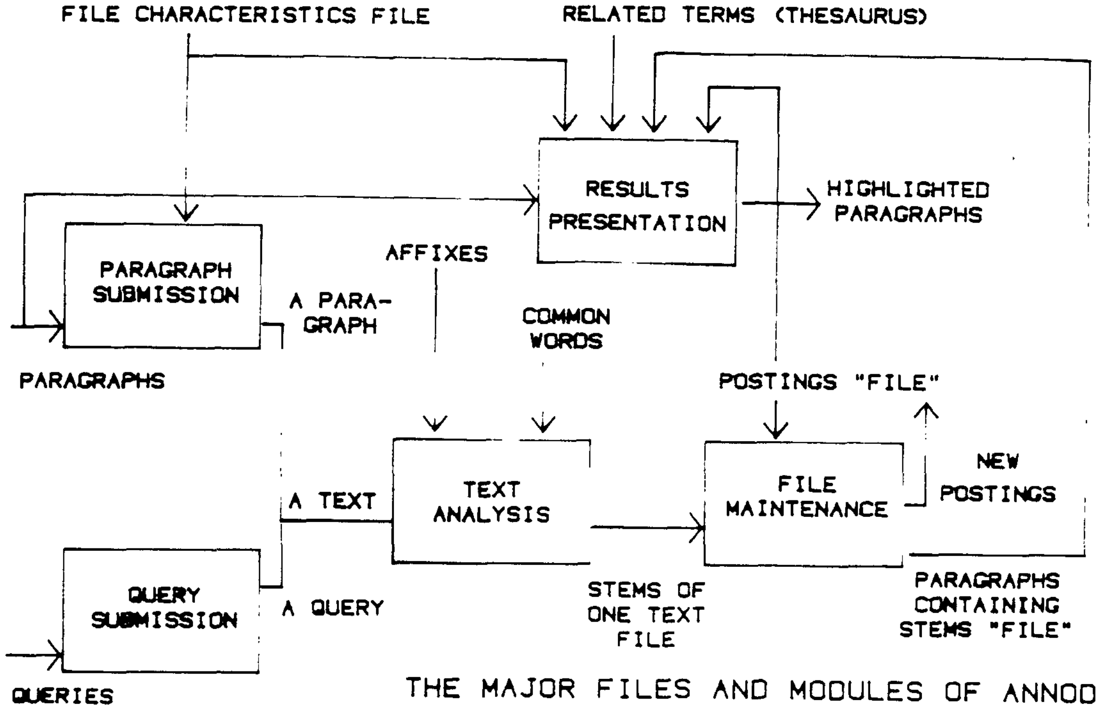
The next step is to process a previously created list of "paragraphs which contain roots" (the postings file) for each root in the expanded query -- "Root XXX appears ...". After the postings for all roots have been processed, those paragraphs having a non-zero similarity are scanned (in hierarchical order) to convert data about roots in common with the query into a rank -- "Ranking the # positively related ...". Paragraphs with poor matches are discarded without being ranked. The most highly ranked ("best") paragraph is automatically displayed, Figure 3c.

B. The Summary Display of Highly Ranked Items

Since human judgement is often better than the machine's, a summary display of the highest ranked paragraphs is provided as shown in Figure 4. In this display, the first line of each paragraph appears along with some numerical data (see below). Based on this display, a user can often locate the material he/she wishes more quickly than by simply displaying the items in the computer's order. The user can select ranked displays limited to synthesis paragraphs (Figure 4b), limited to data paragraphs, or inclusive of all paragraphs (Figure 4a).

The summary display has columns showing (a) the ID of each retrieved paragraph, (b) an indication of the hierarchical position of the paragraph ("H" indicates a heading, "S" indicates a synthesis-statement, "T" indicates a twig, and a blank field indicates a data element [*]), (c) the rank of the item on that line, (d) the percentage match of the paragraph with the query (relative to a hypothetical good match), (e) the actual "points" on which the item is ranked, and, (f) the field of most importance to users, a complete heading or the first line of a leaf paragraph (detail element).

The present algorithm is unable to consistantly rank synthesis statements and data elements. (This detriment could probably be removed by modifications to the similarity measure.) Normally data elements with any significant similarity to a query will outrank very similar synthesis statements. The ranking among synthesis statements appears useable as is the ranking among data elements. The fact that the best synthesis statement in Figure 4b, a near-perfect match to the query, is at rank 202 illustrates the present need to be able to display such paragraphs independently of data elements.



THE MAJOR FILES AND MODULES OF ANNOD

V. THE SIMILARITY MEASURE USED IN ANNOD

A. Overview

Every retrieval system more or less efficiently computes a figure-of-merit which measures the similarity of an item to a query. (E.g., the common boolean based retrieval systems compute a figure-of-merit with only two values: retrieved and not retrieved.) The central principal of ANNOD is to locate and rank those data base paragraphs having the greatest number and variety of the roots in a user's query. The ranking is accomplished by a series of numerical calculations which finally result in the assignment of a single integer value to each paragraph being ranked.

The similarity measure used in ANNOD is empirical in form and is based on the following:

* number of roots in common
* frequency of matching roots in a paragraph (mapped to importance)
* importance of matching roots as noted by the user
* importance of matching roots as noted by a lexicographer
* frequency of matching roots in the file
* length of a paragraph
* size of match
* percentage of paragraph matched
* whether a matching root comes from query or via a thesaurus
* type of paragraph

The specific numbers initially used were chosen based on the author's experience -- and NOT as a result of any computational process. A small number of alterations have been attempted -- seven sets thus far. None of the changes , however, were the result of a systematic study. In every case, a change was made to improve performance for a specific kind of query -- and in particular, for a specific query that did not perform as well as the experimenters thought it should. Thus, the experimenters consider the actual values to be only a starting point for further refinement. Much of the improvement in ANNOD is due to the synergetic interaction of the author, trained in information retrieval, with Dr. Lionel Bernstein, the physician who created most of the data base.

The similarity measure used in ANNOD is defined in Figure 7. There is a non-zero addend for each root, r, present in both the query, q, and a particular paragraph, p. Each addend is a product of three factors: an indicator of the usefulness of the root, ROOTIMP(PAR_r), the importance of the root in the query, $QUERY_{q,r}$, and the importance of the word in paragraph p, PARIMP($PAR_{p,r}$) -- where PAR_r and $PAR_{p,r}$ are the number of times root r appears in the entire collection and in paragraph p, alone, respectively. The remaining addends reflect a bonus, MR($DM_{p,q}$), for having $DM_{p,q}$ distinct roots in common; adjustments, HS_p and DE_p, for being a Synthesis statement, or Detail paragraph; and finally, a negative addend, L(PAR_p), to compensate for the tendency of long paragraphs to have higher scores than shorter paragraphs.

Rev. 9/7/83

$$\text{CROSS}_{p,q} = (\text{ROOTIMP}(\text{PAR}_r) * \text{QUERY}_{q,r} * \text{PARIMP}(\text{PAR}_{p,r}))$$

$$\text{STRENGTH}_{p,q} = \text{CROSS}_{p,q} + \text{MR}(\text{DM}_{p,q}) - \text{L}(\text{PAR}_p) + \text{HS}_p$$

$$\text{PROPORTION}_{p,q} = \text{TN}_{p,q} \;/\; \text{PAR}_p * \text{STRENGTH}_p$$

$$\text{SIMILARITY}_{p,q} = \text{STRENGTH}_{p,q} + \text{PROPORTION}_{p,q} + \text{DE}_p$$

Token Denotes

$\text{CROSS}_{p,q}$ -------- the sum of products of ROOTIMP, QUERY and PARIMP over all distinct roots r

DE_p ------------- if a data element, the STRENGTH of a paragraph which contains every word in the query -- to compensate for the effect of a heading on its synthesis statement

$\text{DM}_{p,q}$ ----------- the number of distinct matching roots in both query and paragraph

HS_p ------------- if a synthesis statement, the SIMILARITY value of its heading

L(f) ------------ the presumed increase in STRENGTH due to length alone, being 3*f

MR(f) ----------- the presumed additional importance of m distinct matching roots, being 1600*f

p ---------------- a specific paragraph

PAR_r ------------ the number of paragraphs which have one or more occurrences of r

PAR_p ------------ the total number of roots (words) in paragraph p

$\text{PAR}_{p,r}$ ---------- the total number of paragraphs which have one or more occurrences of r

PARIMP(f) ------- the presumed importance of a root occurring f times in one paragraph, being a linear interpolation of:

$f=\text{PAR}_{p,r}$	1	2	3	4+
PARIMP(f)	12	18	22	24

$\text{PROPORTION}_{p,q}$ --- the presumed importance of paragraph p due to the PERCENTAGE of its roots also appearing in query q

q ---------------- a specific query

$\text{QUERY}_{q,r}$ -------- the presumed importance of the root r in the query q, being, in general, frequency*18 for the searcher's own words and frequency*12 for thesaurus words

r ---------------- a specific root

ROOTIMP(f) ------ the presumed importance of a root of collection frequency f, being a linear interpolation of:

$f=\text{PAR}_r$	1	10	100	1000+
ROOTIMP(f)	30	20	4	1

$\text{SIMILARITY}_{p,q}$ --- the hypothesized similarity of paragraph p to the query q

$\text{STRENGTH}_{p,q}$ ----- the presumed importance of paragraph p due to the NUMBER of its roots also appearing in query q

$\text{TN}_{p,q}$ ----------- the total number of roots in paragraph p which match roots in the query

FIGURE 7 -- The ANNOD Similarity Measure

VIII. FAILURE ANALYSIS

Some queries fail to retrieve appropriate paragraphs; the most common reasons appear to be:

* an inadequate thesaurus
* an important stem on "deletion" list
* a missing suffix
* inadequacy in the suffixing algorithm
* poor processing of hyphenated words
* lack of pertinent information in the file.

The most common cause of failure has been an inadequate thesaurus. However, the thesaurus in use has not been systematically developed. It is only a demonstration of what can be done. As the thesaurus is systematically developed by content experts, most of the failures attributable to such deficiencies will be eliminated. (An initial effort at thesaurus expansion has increased the thesaurus from about 30 groups of roots to nearly 100. This has eliminated most thesaurus-related failures.) Failures in the suffixing algorithm will be addressed by requiring lexicographer acceptance of every stem (with a sophisticated lexicographer support system, the author has been able to review and, where appropriate, change the suffixing for over 1000 words in an hour.) The quality of the automatic suffixing algorithm increases dramatically once a small portion of text has been processed. For example, mappings like MICE to MOUSE are very important but also very few in number -- most words are regular but many frequently used words are not regular. Thus, new words in queries are likely to be stemmed correctly automatically.

IX. EVALUATION

ANNOD has undergone no comparative evaluation (i.e., comparison of ANNOD to any other system or even to a different version of ANNOD itself). However, there is copious anecdotal evidence that ANNOD's performance is highly cost-effective. The cost of the ANNOD approach is an order of magnitude less than conventional approaches since ANNOD does not require highly-trained, i.e., expensive, human indexing or search intermediaries. A formal evaluation of ANNOD's performance is highly desirable but must await the allocation of adequate resources.

One short evaluation was performed in November, 1982, by Ray Tidman, a medical intern at NLM. He picked ten paragraphs scattered throughout the data base. For each paragraph a short query and a fully expressed query were written which could be answered by the selected paragraph. In every case, the best-ranked detail element or synthesis-statement (as appropriate) was relevant to the query; in every case the "target" paragraph was ranked highly enough to be displayed in the first group of paragraphs in the Summary Display, Figure 6.

Dr. Lionel Bernstein has performed an extensive, albeit non-standard, evaluation of over 60 queries with three judges which documents the likelihood that ANNOD will retrieve useful information in the top few paragraphs [5]. Information needed to answer 85 to 95 percent of the queries was located and displayed in the first few selected paragraphs. Synthesis paragraphs and data paragraphs were shown to be complementary in their information content. ANNOD was successful in locating information in both areas of the text.

Specific evaluation experiments are planned. Recall and precision measurements are planned using other data bases with author-specified relevant items.

[5] Bernstein, Lionel M. and Williamson, Robert E., "Testing of a Natural Language Retrieval System for Full Text Knowledge Bases", a report of the Information Technology Branch, Lister Hill National Center for Bicmedical Communications, National Library of Medicine and JASIS,35:4, p.235-247, July 1984.

X. SUMMARY

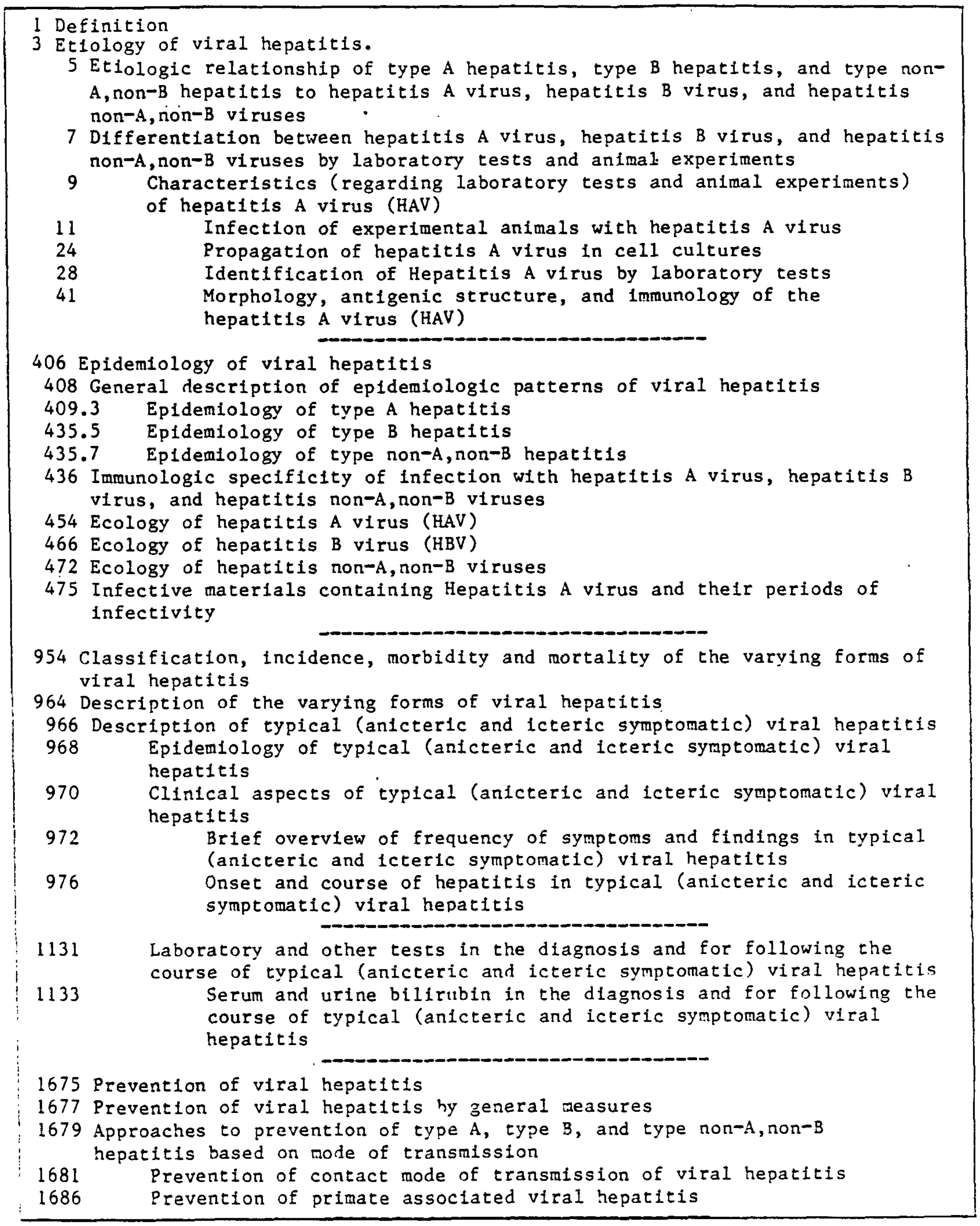
The ANNOD system provides a mechanism by which an end-user can directly access a hierarchically-organized natural-language knowledge base in a fast, effective, user-friendly manner. A user makes a request by entering a natural-language description of the information desired and the system automatically displays that paragraph most similar to the user's description, as well as tokens of several other similar paragraphs for further exploration. Many users of the Hepatitis Knowledge Base have found that the paragraphs retrieved by the present system have satisfied their information needs.

Appendix I - A Brief Overview of Techniques for Accessing Information in ITEMS

Two methods of access were originally supported by ITEMS for the HKB -- both are headings-based. First, all headings could be displayed in physical order -- in essence, a table of contents (Figure A-1). Since this list is over 500 lines long, printed copies were often used. When a suitable heading was located, entry of the heading's ID number caused display of the associated synthesis-statement and one could browse from there (via the commands of ITEMS shown in Table A-1 -- details of which appear in [2]).

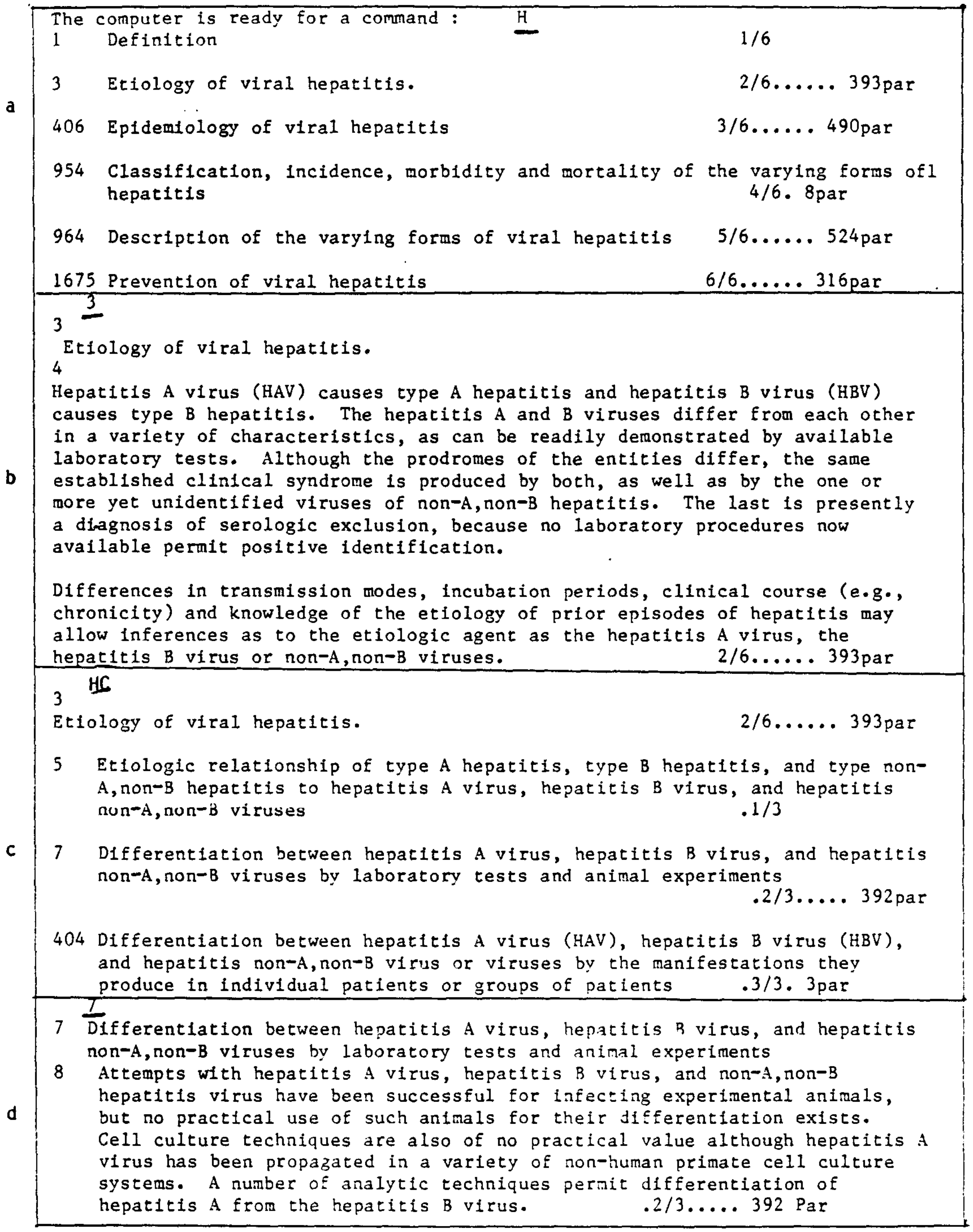
The second approach was to display only the most general (root-level) headings (via the Headings (H) command (Figure A-2a)). The user would select a root and enter its ID number, e.g. "3", to have the associated synthesis-statement displayed (Figure A-2b). The user could then request a display of all headings summarized by the selected heading/synthesis-statement pair by entering the Headings Children (HC) command (Figure A-2c). This "top-down" search strategy was used recursively until either 1) the desired information was obtained or 2) detail elements that could only be read sequentially were reached.

Figure A-3 provides an overview of ITEMS with the portions that implement ANNOD highlighted. ANNOD is basically an alternative method of selecting a paragraph for display. ANNOD's utility lies in its ability to select a single paragraph based on a single user input. Other approaches (e.g., those described above) typically require many decisions of a user and those decisions must be based on marginal, and often inadequate, information.



```
1 Definition
3 Etiology of viral hepatitis.
   5 Etiologic relationship of type A hepatitis, type B hepatitis, and type non-
     A,non-B hepatitis to hepatitis A virus, hepatitis B virus, and hepatitis
     non-A,non-B viruses
   7 Differentiation between hepatitis A virus, hepatitis B virus, and hepatitis
     non-A,non-B viruses by laboratory tests and animal experiments
   9      Characteristics (regarding laboratory tests and animal experiments)
          of hepatitis A virus (HAV)
  11           Infection of experimental animals with hepatitis A virus
  24           Propagation of hepatitis A virus in cell cultures
  28           Identification of Hepatitis A virus by laboratory tests
  41           Morphology, antigenic structure, and immunology of the
               hepatitis A virus (HAV)
                          ----------------------------------

406 Epidemiology of viral hepatitis
 408 General description of epidemiologic patterns of viral hepatitis
 409.3    Epidemiology of type A hepatitis
 435.5    Epidemiology of type B hepatitis
 435.7    Epidemiology of type non-A,non-B hepatitis
 436 Immunologic specificity of infection with hepatitis A virus, hepatitis B
     virus, and hepatitis non-A,non-B viruses
 454 Ecology of hepatitis A virus (HAV)
 466 Ecology of hepatitis B virus (HBV)
 472 Ecology of hepatitis non-A,non-B viruses
 475 Infective materials containing Hepatitis A virus and their periods of
     infectivity
                          ----------------------------------

954 Classification, incidence, morbidity and mortality of the varying forms of
    viral hepatitis
964 Description of the varying forms of viral hepatitis
 966 Description of typical (anicteric and icteric symptomatic) viral hepatitis
 968      Epidemiology of typical (anicteric and icteric symptomatic) viral
          hepatitis
 970      Clinical aspects of typical (anicteric and icteric symptomatic) viral
          hepatitis
 972           Brief overview of frequency of symptoms and findings in typical
               (anicteric and icteric symptomatic) viral hepatitis
 976           Onset and course of hepatitis in typical (anicteric and icteric
               symptomatic) viral hepatitis
                          ----------------------------------

 1131     Laboratory and other tests in the diagnosis and for following the
          course of typical (anicteric and icteric symptomatic) viral hepatitis
 1133          Serum and urine bilirubin in the diagnosis and for following the
               course of typical (anicteric and icteric symptomatic) viral
               hepatitis
                          ----------------------------------

 1675 Prevention of viral hepatitis
 1677 Prevention of viral hepatitis by general measures
 1679 Approaches to prevention of type A, type B, and type non-A,non-B
      hepatitis based on mode of transmission
 1681     Prevention of contact mode of transmission of viral hepatitis
 1686     Prevention of primate associated viral hepatitis
```

Figure A1. Portions of the HKB Table of Contents



```
The computer is ready for a command :      H
1    Definition                                              1/6

3    Etiology of viral hepatitis.                            2/6...... 393par

406  Epidemiology of viral hepatitis                        3/6...... 490par

954  Classification, incidence, morbidity and mortality of the varying forms ofl
     hepatitis                                                4/6. 8par

964  Description of the varying forms of viral hepatitis    5/6...... 524par

1675 Prevention of viral hepatitis                         6/6...... 316par
```

```
   3
3
 Etiology of viral hepatitis.
4
Hepatitis A virus (HAV) causes type A hepatitis and hepatitis B virus (HBV)
causes type B hepatitis.  The hepatitis A and B viruses differ from each other
in a variety of characteristics, as can be readily demonstrated by available
laboratory tests.  Although the prodromes of the entities differ, the same
established clinical syndrome is produced by both, as well as by the one or
more yet unidentified viruses of non-A,non-B hepatitis.  The last is presently
a diagnosis of serologic exclusion, because no laboratory procedures now
available permit positive identification.

Differences in transmission modes, incubation periods, clinical course (e.g.,
chronicity) and knowledge of the etiology of prior episodes of hepatitis may
allow inferences as to the etiologic agent as the hepatitis A virus, the
hepatitis B virus or non-A,non-B viruses.                  2/6...... 393par
```

```
   HC
3
Etiology of viral hepatitis.                                 2/6...... 393par

5   Etiologic relationship of type A hepatitis, type B hepatitis, and type non-
    A,non-B hepatitis to hepatitis A virus, hepatitis B virus, and hepatitis
    non-A,non-B viruses                                       .1/3

7   Differentiation between hepatitis A virus, hepatitis B virus, and hepatitis
    non-A,non-B viruses by laboratory tests and animal experiments
                                                                .2/3..... 392par

404 Differentiation between hepatitis A virus (HAV), hepatitis B virus (HBV),
    and hepatitis non-A,non-B virus or viruses by the manifestations they
    produce in individual patients or groups of patients      .3/3. 3par
```

```
   7
7  Differentiation between hepatitis A virus, hepatitis B virus, and hepatitis
   non-A,non-B viruses by laboratory tests and animal experiments
8   Attempts with hepatitis A virus, hepatitis B virus, and non-A,non-B
    hepatitis virus have been successful for infecting experimental animals,
    but no practical use of such animals for their differentiation exists.
    Cell culture techniques are also of no practical value although hepatitis A
    virus has been propagated in a variety of non-human primate cell culture
    systems.  A number of analytic techniques permit differentiation of
    hepatitis A from the hepatitis B virus.          .2/3..... 392 Par
```

Figure A2   Top-Down Searching of Heirarchically Organized Information

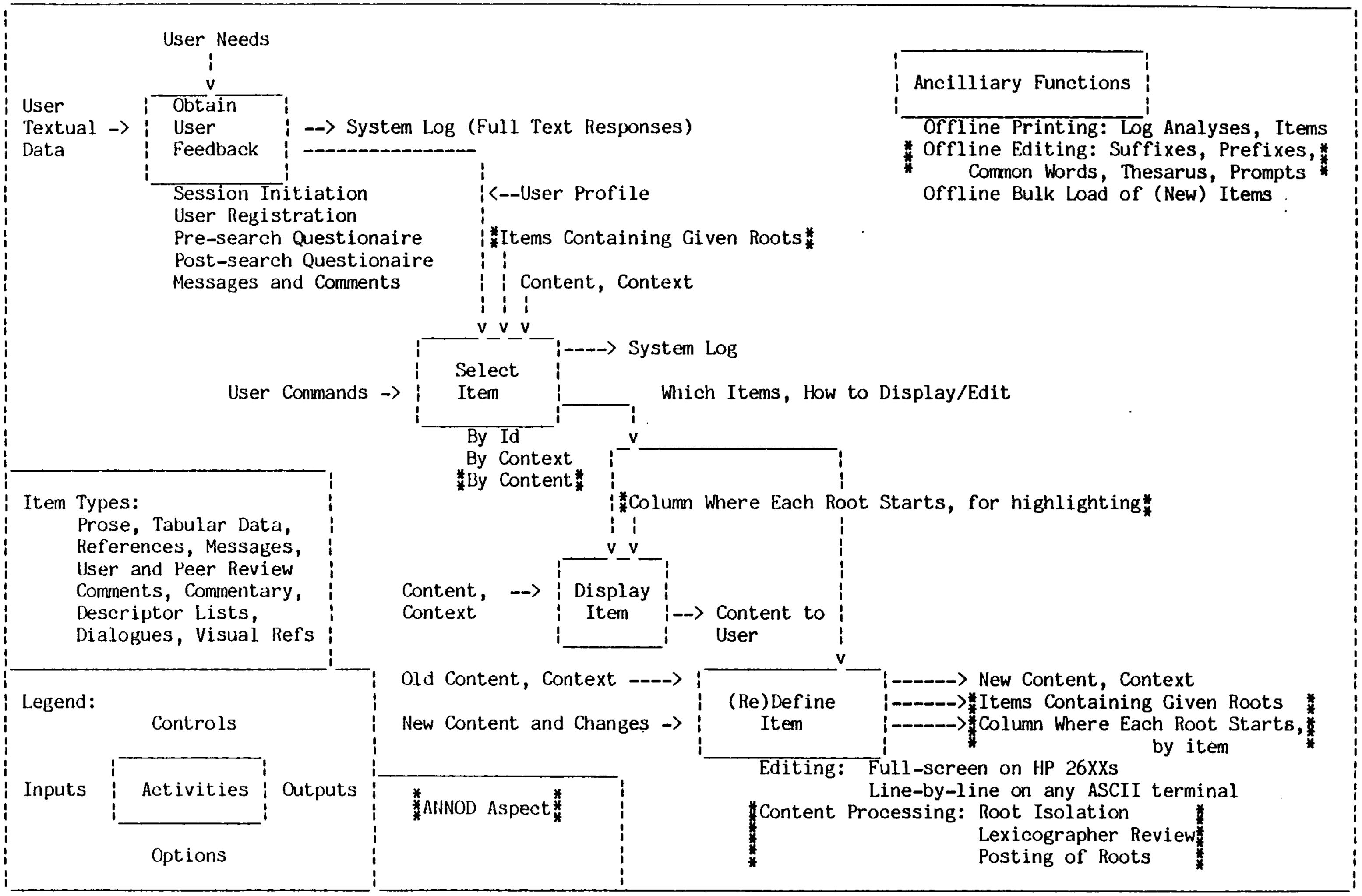


Figure A3. Major Processes Supported by ITEMS -- An Interactive TExt Management System