ELECTRONIC **WORKSHOPS IN COMPUTING**

Series edited by Professor C.J. van Rijsbergen

Jonathan Furner, School of Information and Media Studies, and David Harper, School of Computer and Mathematical Studies, The Robert Gordon University, Aberdeen, Scotland. (Eds)

Information Retrieval Research

Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research, Aberdeen, Scotland, 8-9 April 1997

Paper:

Text-Level Structure of Research Papers: Implications for Text-Based Information Processing Systems

N. Kando

Published in collaboration with the
British Computer Society

BCS

Springer

Text-level Structure of Research Papers : Implications for Text-Based Information Processing Systems

Noriko Kando

Research and Development Department, National Center for Science Information Systems (NACSIS)
Tokyo, Japan
who is now in Department of IR Theory, Royal School of Librarianship
Copenhagen, Denmark

Abstract

This paper discusses the implication of text-level structure for text-based information processing systems. In this paper, text-level structure of research papers is described with a set of typical functional components of research papers such as, background, purpose, methods, etc. and their order in a text. In order to suggest various applications, the experiments of retrieval and passage extraction were conducted using a manually structure-tagged fulltext database of research papers. As a result, we show that searching full-length texts using text-level structure achieved higher precision, compared to the searching without it. The paper also shows examples of extracted passages and suggests the application of text-level structure for text-based information systems, including passage extraction, browsing, and navigation within/across texts.

1 Introduction

Text is not a mere collection of sentences, but is the set of concepts carefully constructed by the author in order to convey the message effectively. Each sentence in a text is interpreted in relation to the other sentences in the text. Although current information retrieval systems are usually focused on terms or sentence-level structures, the text- level phenomena such as, global structure of the text, relation among the passages within a text, anaphoric references, etc. and characteristic aspects of the text type should be considered in the design of text-based information processing systems.

The phenomena within units larger than a sentence have been investigated in discourse linguistics and text linguistics. Among the various approaches to text-level structure, this paper focuses on the typical structure of the genre or the text type. It is known that each text type has a typical structure. Many researchers have shown that the research papers and their abstracts possess a highly typical structure [1-5]. Such a typical structure can be described with a set of typical components of the text type and their order in a text[2]. “Background”, “reference to the previous research”, “purpose”, “methods”, “results”, “discussion”, and “conclusion” are examples of typical components of the text type of research papers. The set of such components is somewhat predictable in each text type. It is sometimes mentioned as the functional structure of the text since each component often represents the function or role that each part of the text plays in the text. It is a natural structure of informational content of the text which is familiar with the people who use the text type to communicate as a kind of social convention, therefore it is expected to be usable by users in a search process [6].

Text-level structure of research papers has been applied to various areas such as: indexing [7], automatic abstracting [8-11], reference interviews [12-13], text retrieval [4,6,14], browsing in a text [15], message understanding, and the design of user interfaces in electronic journals [16-18]. My colleagues and I have conducted a series of analyses on text-level structures of several kinds of texts (or more precisely, information objects), for example, we have delineated the typical structure of Japanese research papers [19-20], English and Japanese newspaper articles [21-23], Nursing records [24], pictures and videotapes of TV news [25], conducted automatic detection of the structure [26-28], and suggested various applications [24].

For the application for information retrieval, a search using text-level structure is expected to be more effective than one that does not. This improvement is not only by detecting the central theme in a text, but also by distinguishing the role or function each concept plays in the text or the relationship among them [4,7].

In addition to this, text-level structure of research papers is suggested as a model of the scientific research process [17], and a suitable frame for interacting with a user on the matter of the user’s situation in the research process [6]. Some of the typical components of research papers also consist of a part of the criteria of judging relevance or appropriateness of retrieved documents to users’ tasks or problems [29-30]. Users’ situations, problems or tasks are integral components to information needs [31-32]. The situation of each user is more complex than the text-level structure represents, but this can be a small step towards deriving factors that may relate to it from texts, and incorporating these factors to information systems.

Moreover, users' situations cans be changed dynamically during the interaction with information systems [33-34]. Text-level structure enables a more flexible interaction by providing various ways of displaying, browsing, or navigating within/among the texts through links other than topical or semantic relationships or logical structure of document style. It may facilitate users to obtain insight into the content of databases through interaction.

1.1 Research Goals

Based on previous research, in order to suggest various applications of text-level structure, this paper experiments on a pilot system of fulltext database of Japanese research papers using text-level structure, including full-length text retrieval in both with ranked-output and without it, passage extraction, and comparison of extracted passages. The paper also discusses the implication of text-level structure analysis in text-based information processing systems.

Passage retrieval is convenient for users who wish to obtain just a passage immediately relevant to their needs, rather than the long text which includes it. It is also reported to improve both precision and recall [35]. But a passage is not an independent entity like a whole text. It should be interpreted and examined for validity in relation to the rest of the text. Text-level structure can be used to describe the relations among the passages within the text.

Readers examine the texts or passages, then compare and integrate their contents. Based on this, readers do information works like writing papers, decision making, problem solving, etc. Comparison of texts or passages may lead to the discovery of a new relationship among them [36]. This paper is especially interested in the application which may support human information works by comparing extracted passages from different texts.

Comparison or summarization of multiple texts based on a knowledge based system [37] or using message understanding technique [38] are reported. However, the applications suggested here are based on "light" text processing which keeps the text in its original form and facilitates users' strategic or analytical reading or searching of texts by providing flexible ways of display and navigation by text-level structure. Moreover, satisfactory performance of the knowledge based approach can only be obtained when the domain is restricted to a narrow subject area. The approach of this paper is limited to the genre of research papers but no restriction on subject area is pre-supposed.

The next section describes the methods of analyzing the text-level structure, and section 3 describes the methods and results of experiments. Section 4 discusses the implication of text-level structure for text-based information processing systems and suggests future studies.

2 Analysis of Text-Level Structure of Research Papers

Text-level structure, as discussed above, is analyzed using a set of analytical categories which represent the typical components of the text type. The structure of each text is described by occurrence of categories and their order in the text.

2.1 The Categories : Typical Components of Research Papers

The set of Categories is presented in Figure 1. It was prepared through content analysis of 40 writing manuals and 127 Japanese research papers selected from four disciplines of medicine, physics, economics, Japanese literature, and interviews with researchers [19-20]. It has been modified through applying to samples from other disciplines [26-27,39-41] and the inter-coder consistency tests [41]. Each Category represents the role or function that the part of text plays in the text, the relationship between a part of text and the whole text, and between parts in the text and other parts in other text.

They are arranged hierarchically. The most specific level of Categories are assigned to each sentence. They can be translated into an upper level via hierarchy links when the overall structure of the text is studied. The characteristics and definition of upper level Categories are inherited to specific ones through the hierarchy links. This inheritance was also used in the rules for automatic detection of the Categories.

The Categories were assigned to all the sentences in the sample. That is, the informational content of research papers from various disciplines is described with the same repertory of the Categories, which can be used as a common framework for them. Text-level structure in each research paper was described with the Category occurrence and order in it. It was more complex than the "IMRD", or "Introduction, Methods, Results, and Discussion" organization, which is frequently mentioned as the structure of research papers in writing manuals [42] and roughly reflected in the sectional headings of empirical research papers. But these patterns

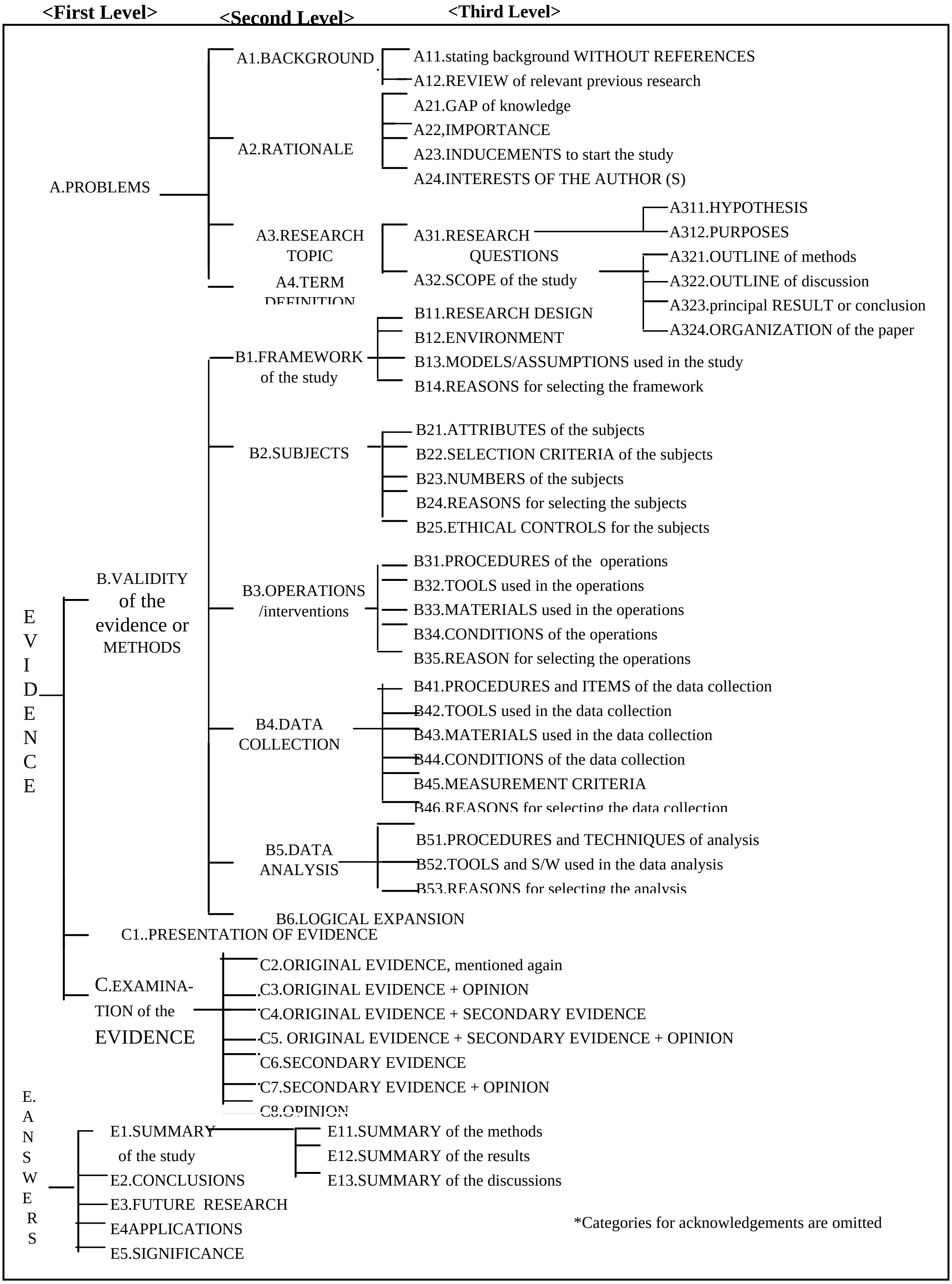

<First Level>
<Second Level>
<Third Level>
A.PROBLEMS
A1.BACKGROUND
A11.stating background WITHOUT REFERENCES
A12.REVIEW of relevant previous research
A2.RATIONALE
A21.GAP of knowledge
A22,IMPORTANCE
A23.INDUCEMENTS to start the study
A24.INTERESTS OF THE AUTHOR (S)
A3.RESEARCH TOPIC
A31.RESEARCH QUESTIONS
A311.HYPOTHESIS
A312.PURPOSES
A32.SCOPE of the study
A321.OUTLINE of methods
A322.OUTLINE of discussion
A323.principal RESULT or conclusion
A324.ORGANIZATION of the paper
A4.TERM DEFINITION
EVIDENCE
B.VALIDITY of the evidence or METHODS
B1.FRAMEWORK of the study
B11.RESEARCH DESIGN
B12.ENVIRONMENT
B13.MODELS/ASSUMPTIONS used in the study
B14.REASONS for selecting the framework
B2.SUBJECTS
B21.ATTRIBUTES of the subjects
B22.SELECTION CRITERIA of the subjects
B23.NUMBERS of the subjects
B24.REASONS for selecting the subjects
B25.ETHICAL CONTROLS for the subjects
B3.OPERATIONS /interventions
B31.PROCEDURES of the operations
B32.TOOLS used in the operations
B33.MATERIALS used in the operations
B34.CONDITIONS of the operations
B35.REASON for selecting the operations
B4.DATA COLLECTION
B41.PROCEDURES and ITEMS of the data collection
B42.TOOLS used in the data collection
B43.MATERIALS used in the data collection
B44.CONDITIONS of the data collection
B45.MEASUREMENT CRITERIA
B46.REASONS for selecting the data collection
B5.DATA ANALYSIS
B51.PROCEDURES and TECHNIQUES of analysis
B52.TOOLS and S/W used in the data analysis
B53.REASONS for selecting the analysis
B6.LOGICAL EXPANSION
C1..PRESENTATION OF EVIDENCE
C.EXAMINA-TION of the EVIDENCE
C2.ORIGINAL EVIDENCE, mentioned again
C3.ORIGINAL EVIDENCE + OPINION
C4.ORIGINAL EVIDENCE + SECONDARY EVIDENCE
C5. ORIGINAL EVIDENCE + SECONDARY EVIDENCE + OPINION
C6.SECONDARY EVIDENCE
C7.SECONDARY EVIDENCE + OPINION
C8.OPINION
E. ANSWERS
E1.SUMMARY of the study
E11.SUMMARY of the methods
E12.SUMMARY of the results
E13.SUMMARY of the discussions
E2.CONCLUSIONS
E3.FUTURE RESEARCH
E4APPLICATIONS
E5.SIGNIFICANCE
*Categories for acknowledgements are omitted

Figure 1. The Categories

were eventually categorized and described by the combination of “basic patterns”, “sub-patterns”, repetition or omission of “sub-patterns”, and “exception patterns”. The detailed list of the patterns and number of occurrences in each level of Categories were reported [19-20].

A relatively small number of indicative clue phrases for each Category were revealed through the analysis. These phrases and the patterns of Categories were used in the rules for automatic Category detection.

Text structure is one of the fundamental characteristics of informational contents of texts, and can be applied to various processing methods such as: information retrieval, automatic abstracting, and information extraction. However, manual assignment of the Categories is rather time-consuming and is a labour intensive task, therefore making it impractical in operational settings. As a premise of application, automatic detection of Categories is necessary.

2.2 Automatic Detection of the Categories

Feasibility of automatic Categories detection was revealed through feasibility studies [24,26-28]. Three kinds of rules, *i.e.*, (a) Indicative clue phrases, (b) Category order, and (c) Category scope, were used in these studies. The rules using indicative clue phrases are essentially templates. They may include groups of indicative clues or phrases, a Category of the previous sentence, and citations. The rules for Category scope identify the scope that a Category continues in the text. These are mainly based on conjunctions and anaphoric expression.

Some rules provide stronger evidence than others, and each rule is assigned a weight. The score for weighting is initially calculated as the probability that the sentence falls into the specific Category when it is matched to the premise of the rule in the training corpus, and has been adjusted during successive trials with it.

The results of these studies indicate the feasibility of automatic assignment of Categories using surface level natural language processing. Based on this, the next section describes the results of the experiments using a pilot system with structure tagged fulltext database, and suggests various applications for text-based information processing systems.

3 Experiment

3.1 Database and Search Engine

The experimental fulltext database consisted of fifty Japanese research papers on Viral Hepatitis type C, which was one of the corpora used in automatic Categories detection [26]. They were selected systematically from an operational large-scale medical bibliographic/abstract database with conditions of “published in 1991”, “containing the term ‘Viral Hepatitis type C’ or ‘non A non B Viral Hepatitis’”, “original papers”, and “written in Japanese”. The fulltext of papers was OCRed after gaining permission from the publishers.

Tags for the Categories, which represent the typical functional components of research papers that is shown in Figure 1, and the components of logical structure of documents, *i.e,* <article>, <title>, <sec> for section, <p> for paragraph, etc., were assigned manually. An example of database records is shown in Figure 2.

```
<article><title>Interferon Therapy of Non A Non B
Hepatitis</title>
<body><sec><h1>Introduction</h1>
<p><s><A11>It has been reported that interferon (IFN)
treatment declines the serum transaminase level and
improves the histological condition.</A11></s>
<s><A21>However there are some problems in the methods of
administration since many cases whose level of serum
transaminase increased after continuous administration of
IFN for four weeks were reported.</A21></s></p>
<p><s><A312>In order to discuss the more effective method
of IFN administration, The authors conducted the survey
```

Tags represents the components of logical structure of documents, such as: <article> for the article as a whole, <title> for title, <abs> for abstract, <body> for body of text, <sec> for section, <h1> for section heading, <p> for paragraph, <s> for sentence, <cap> for caption for figure, and the Categories shown in Figure 1 such as: <A11>, <A21>,<B21> and so on.

Figure 2. English Translation of an Example of the Experimental Fulltext

The search engine is OpenText 6 (OpenText Corp., Canada), which is a fast text searching system. In this system, a database is seen as one long string and the queries are based on *sistring*, *i.e*. a semi-infinite string that starts at that position and extends arbitrarily far to the right, or to the end of the text [43]. The part of text enclosed by a beginning tag <tag> and an ending tag </tag>, which are specified in the data dictionary, is called a *region*. It is a unit for a search and results set. “Including” and “within” operations among regions are available. With this engine, complex queries with structure relationship can be processed, for example, “any articles in which the term ‘rat’ occurred in the Category of ‘B21. Attribute of subject’ ”, or “any paragraphs in which word ‘weight’ occurred in ‘C1. original evidence of the article’ in the articles retrieved by this query ” , and so on.

The search with statistical ranked-output uses OpenText’s “RankMode Relevance1”, which “ranks the members of the returned set based on the term frequency, the document length, and the total number of all words” [44].

3.2 Experiments 1: Full-Length Text Searching

3.2.1 Procedure

The purpose of this experiment is to test the effectiveness of text-level structure in fulltext searching both with statistical ranked-output and without it.

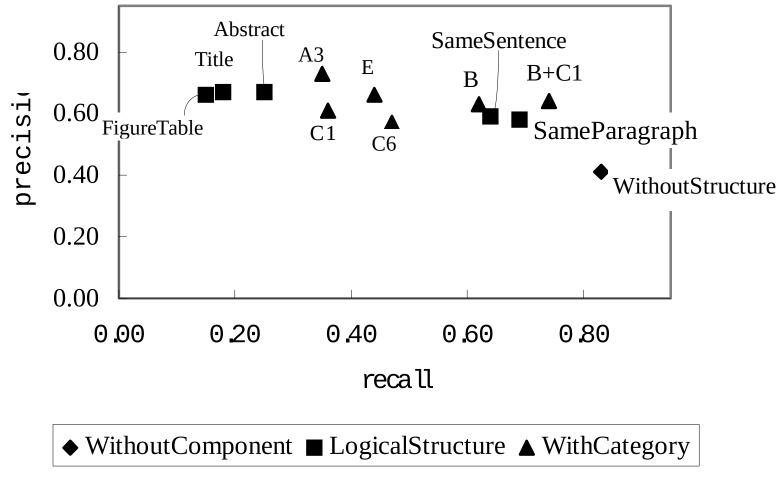
The searches without ranked-output were conducted with these strategies, (a) without any specification of Categories nor logical structure, (b) using components of logical document structure (title, abstracts, same paragraph and same sentence, and title/caption of table/figure), and (c) using the Categories. In strategy (c), the Categories were combined with higher level ones as shown in Figure 3. Strategy (a) is based on co-occurrence of search terms in a text and the others are based on term occurrence in specific Categories/components in a text.

The searches with ranked-output were conducted with two strategies, *i.e.* (a) without any specification of Categories nor logical structure and (b) ranking based on term occurrence in the Categories “B. validity and Method” and “C1. Evidences”, which were effective in searching without ranked-output, and ranking based on combination of term occurrence in these Categories and the whole text.

Sixteen search topics were collected from medical researchers for the experiment. Nine topics were omitted since the number of relevant documents for them was less than four or more than 25% of the entire database. Search statements were constructed manually including synonyms.

3.2.2 Results

The results for the search without ranked-output were shown in Figure 3. In general, precision was increased when Categories or components were specified. “A3. Research Topic” are the parts of the text in which usually the central concepts of the texts, or purposes of research reported in the articles, are stated. Specifying this Category increased precision acutely but the recall ratio dropped. “B. Methods and C1. Evidence” were effective in increasing precision while keeping recall level. By specifying the second level Categories of “B1”,“B2”, “B3”, “B4”, and “B5” instead of “B”, precision increased without significant decrease of recall.



For Categories: A3: Research Topic, B: Methods, C1: Evidences, C6: Secondary Evidences, E: Answer / Summary

Figure 3: Results of Full-Length Text Search without Ranked-output (Average)

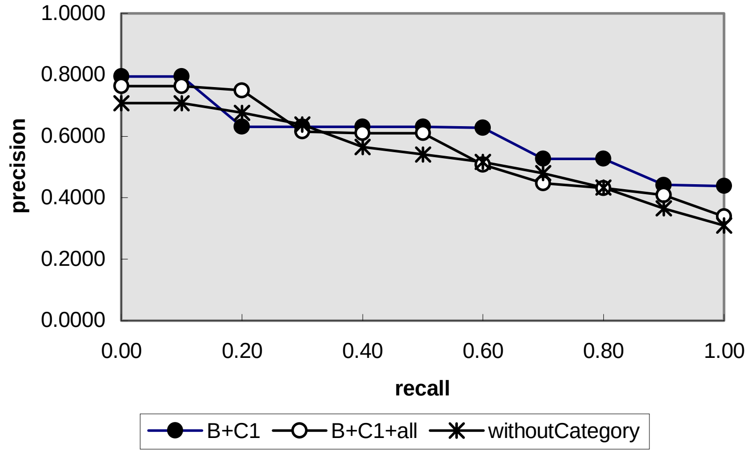


Figure 4: The Result of Full-Length Text Search with Statistical Ranked-output

The results of the search with statistical ranked-output were shown in Figure 4. Searches with Categories “B.Methods” and “C1.Evidence”, and those combinations with occurrence in the whole text were more effective than those that did not use any Category. The effects of the Categories differed in each search topic.

A specific combination of one of the detailed Categories and each term in a query is expected to increase the discrimination and precision of search because most of the concepts in the search topics are related to the specific Categories under “B. Methods” or “C1. Evidence”. For example, when a combination of term ¦ Categories is represented in parentheses, one of the search topics can be represented as “ (the difference in the effect ¦ B41. Items measured or C1.Evidence) of (interferon therapy ¦ B31.Procedure of Intervention) of (viral hepatitis type C ¦ B21. Attributes of Subjects) with (sex ¦ B21. Attributes of Subjects) ”. Selecting appropriate Categories for each search term required expertise. Further investigation is needed on the automatic formulation of the search statement from users’ input.

One of the medical researchers who provided search topics said,

> “…I’m interested in the papers on interferon therapy of Viral Hepatitis which uses Davis’ criteria for evaluating the prognosis status, because I’m currently writing a paper which uses this criteria, so I’d like to know any other papers that use the same criteria and compare the results. ..”

And he said,

> “..... Before our project team had decided the criteria for prognosis evaluation, we collected the papers on interferon therapy of Viral Hepatitis as much as we could, then scanned just the parts of methods used for evaluating the prognosis. And if a candidate was found in a paper, we examined whether the setting reported in the paper was appropriate for ours or not,”

This comment is an example of the relevance judgment and usage of texts, differing according to the users situation. The Categories discussed here can be used both for searching texts stating a specific measurement criteria like “prognosis evaluation criteria”, and displaying specific parts like “methods” of retrieved texts.

3.3 Experiments 2: Passage Extraction

The set of Categories can be used as a framework to extract passages with a specific role or function in a text. It facilitates analytical comparison of informational contents across texts and provides the possibility for users to get insight about the content of databases, hence supporting users’ information work like decision making or problem solving. Sentences with a specific Category or with a combination of Categories and terms can be extracted from the retrieved set or the database. For example, comparison of passages described “methodology” or “term definition” across the texts in the retrieved set is supposed to be effective in comparing each papers’ standpoints and trends among them.

The examples were shown in Appendix 1 and 2. In order to keep cohesion within each extracted passage, the previous sentence was extracted together when the sentence to be extracted had a demonstrative pronoun or indefinitive pronoun as the first words of sentences, next to the first word of conjunction, or when the previous sentence was a topic sentence of the paragraph.

3.3.1 Extract passage stating "facts"

Appendix 1(a) showed the extracted passages which stated the difference in the effectiveness of interferon therapy of VHC in age. Since the query asks facts about it, the sentences are extracted by matching the term in query and in sentences with Category "C1. Evidence". Thirteen sentences and five tables were extracted from five papers. These were small enough to read through and easy to compare. Without using Categories, the number of extracted sentences was much higher.

Extracted sentences from two papers said that there was no significant difference on effectiveness of interferon therapy in age, while the sentences from three other papers said that there were significant differences in age. These results showed that this question had not been resolved yet at the time when the database was constructed, and suggests that a detailed context of each passage included should be examined.

Appendix 1(b) showed that the extracted passages about "Hepatitis C Virus(HCV) positive rate in blood donors". As a result, no sentences extracted from "C1. Evidence". Six sentences and one table from five papers were extracted from "C6. Secondary Evidence". These were small enough to read through and indicated that the HCV positive rate of Japanese blood donors was about 1-2%.

3.3.2 Extracted "State-of-the-Art" Sentences

Appendix 2 shows examples of extracted passages with "A12. Review of relevant previous research" and "A21. Gap of knowledge" from the entire database. Total number of sentences in them was 143. These provide the information about the state-of-the-art of VHC research. The extracted set of passages is an aggregation of small reviews of state-of-art from the each author's point of view. The content might not have sufficient specificity for specialists of this topic, but can be a good background overview for newcomers to the area or non-specialists. Writing a review article demands a lot of time and expertise, but with the Categories, users can extract from a database, passages which may be helpful to know the background situation. Moreover, a common tendency was found in the documents cited in these extracted passages and their order. This suggested that each extracted passage shares, to some extent, a common informational content with others.

These examples showed the possibility of an alternative way of fulltext database with the Categories; extracting related passage across texts. It provides the possibility for users to examine the content of databases and to find new relationships among them. But each extracted passage is not an independent entity and should be interpreted and examined for its validity in the context of the original text. Browsing within the text, guided by the links of Categories or logical structure, is available as described in the next section.

3.3.3 Browsing within a Text

To know the context of each extracted sentence; any unit of text, such as a paragraph, section , or specific Categories related to each sentence, can be displayed using the relationships specified by tags. Browsing using logical structure or Categories is available. For example, sentences with Categories "A312. Purpose", "B1. Research Design", "B2. Subject", "B3. Intervention", "B4. Measurement", and "B5. Statistical Analysis" are supposed to be essential parts in examining the validity of evidence reported in the text. This is also helpful for relevance judgment of retrieved documents .

3.3.4 Other Application : Classification of references

The set of Categories can be used to classify the references cited in each research paper according to the Category of each sentence which cites the reference. The total number of citations made in the experimental databases were 703. Although most of them were cited in the sentences of the Categories "B6. Secondary Evidence"(55.1%) or "A12. Relevant previous research"(27.1%), citation was also made in the sentences of Categories like "B45. Measurement Criteria", "B46. Reasons for selecting the data collection", "A24. Interests of Authors", "B24. Reasons for selecting the subjects", and "B31. Procedures of the Operations".

The list of references cited in each paper is one of the important sources for users to identify the authors' standpoints, and can be used as a search key for related documents. Distinguishing the context of each citation by the Categories will be effective in both cases.

4 Discussion

4.1 Implication of Text-Level Structure for Text-Based Information Processing Systems

The size of the database is one of the fundamental factors we should consider in the effectiveness of information retrieval. Since the size of the database used in this study was small, even though the records in it were systematically selected from an operational large scale database, we can not derive concrete conclusion regarding search effectiveness. However, the results of the experiments indicated some promising lines of investigation relating to text-level structure.

Using text-level structure, a flexible way of display and interaction with the system, such as passage extraction, browsing and navigation within/among texts, comparison of passages extracted across texts is available. These are helpful for users to get insight into the content of database or retrieved records.

The Categories proposed by this series of research has possibility to enhance the interaction between the user and the system by providing links other than topic and semantic or logical structural to the components within the text or in other texts. For example, during the browsing a text, a user found that the particular passage was interesting for him/her. A search of similar texts or passages can be conducted with a function of the search engine, based on terms and Categories occurring in the passage. This makes a search process as a continuous browsing and searching in a highly interactive setting.

Another possible application is the approach for analyzing text-level structure and the concepts of Categories using elaborated templates. It is expected to be applicable to information extraction. Information extraction is one of the fundamental and promising techniques for automatic abstracting, text summarization, integration, question, and answering .

Some comments from the participant suggest the relationship between text-level structure and users' situations in the research process. Further research is needed on this.

4.2 Further Studies

The next step of the studies will be experiments using a large scale database with users. To utilize all the possible search functions in an interactive setting, design of the graphical user interface is one of the crucial points. In the process of developing the user interface, level of Categories used for search, unit of display, methods of query formation using the Categories, and the way of exposure of the Categories to users, should be decided based on users' cognition and search effectiveness. An evaluation method for such a system also requires investigation. Applicability for text other than Japanese also needs to be investigated.

References

1. Samuels, SJ,et al. Adults' use of text structure in the recall of a scientific journal article. J of Educ Res. 1988;18:171-174

2. van Dijk, TA. Macrostructures: an Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition. Hillsdale, Lawarence Erlbaum Assoc., 1980, 317p.

3. Crookes, G. Towards a validated analysis of scientific text structure. Applied Linguistics. 1986;l.7:57-70

4. Liddy, ED. The Discourse level structure of empirical abstracts ; an Exploratory Study. Inf Proc and Manag. 1991;27:55-81

5. Kircz, JG. Rhetorical structure of scientific articles; the case for argumentational analysis in information retrieval. J Doc.1991; 47:354-372

6. Oddy, RN, Liddy, ED, Balakrichnan B, Bishop, A, Elewononi, J. Martin, E. Towards the use of situational information in information retrieval. J Doc. 1992;48:123-171

7. Fuller, SS. Schema theory in the representation and analysis of text. Ph.D. theses, Univ. Southern California, 1984. 189p. available from U.M.I. Order No.DA8500206

8. Paice, CD. Constructing literature abstracts by computer : techniques and prospects. Inf Proc Manag. 1990;.26:171-186

9. Paice, CD; Jones, P. The Identification of important concepts in highly structured technical papers. Proceeding of 16th ACM/SIGIR. 1993, pp 69-78.

10. Johnson, FC, Paice, CD, Black, WJ, Neal, AP The application of linguistic processing to automatic abstract generation. J Doc Text Manag. 1993;1: 215-241.

11. Endres-Neggemeyer, B, et al. How to implement a naturalistic model of abstracting : four core working steps of an expert abstractor. Inf Proc Manag. 1995;31: 631-374.

12. Allen B. Text structures and the user-intermediary interaction. RQ. 1988;27:535-541

13. Allen, B. Recall Cues in known-item retrieval. J Amer Soc Inf Sci. 1989;40:246-252

14. Liddy, E. Myang, SH, DR-LINK. SIG-IR Forum. 1994;18;1-20

15. Miike, S, Itoh, E, Ono, K, Sumita, K. A Full-text retrieval system with a dynamic abstract generation function. Proceedings of the 17th ACM-SIGIR. 1994, p.152-161.

16. Dillon, A. Readers' models of text structures : the case of academic articles. Int J Man-Machin Stud. 1991;35:913-925.

17. Dillon, A.Designing usable electronic text : ergonomic aspects of human information usage. Taylor & Francis. 1994, 195p.

18. Dillon, A, Schaap, D. Expertise and the perception of shape in information. J Amer Soc Inf Sci. (in press)

19. Kando, N Structure analysis of information media using Categories : structure of research articles. Tokyo, Keio University, Unpublished Master thesis, 1991, 227p. (in Japanese)

20. Kando, N. Structure of research articles. IPSJ SIG Notes (92-FI-25). 1992;92:39-46 (in Japanese, with English abstract)

21. Ueda, S, Kando, N, Koshizuka, M. A prototype of question-answering system based on analysis and synthesis of news paper text. IPSJ SIG Notes, 1992 (in Japanese)

22. Kando, N. Structure of news stories : As relating to the indexing and retrieval. Journal of Japan Indexers Association. 1995;19:1-17 (in Japanese)

23. Kando, N. Text structure analysis based on human recognition : cases of Japanese newspaper articles and English newspaper articles. Bulletin of the National Center for Science Information Systems. 1996;8: 107-129 (in Japanese, with English abstract)

24. Kando, N. Text Structure of information media: as a framework for content analysis. Tokyo, Keio University. Unpublished Ph.D. Thesis. 1995 (in Japanese)

25. Ueda, S, Koshizuka, M, Kando, N. Framework for image recognition and alternative indexing method for image. IPSJ SIG Notes (95-HC-28). 1995; 96:55-60 (in Japanese, with English abstract)

26. Kando, N. Functional structure analysis of the research articles and Its application.. Annals of Japan Society of Library Science. 1994;40:49-61 (in Japanese, with English abstract)

27. Kando, N. Functional structure analysis of research articles selected from three specialties : automatic Category assignment. Library and Information Science, 1994;31:.25-38 (in Japanese, with English abstract)

28. Kando, N. Structure of research articles with various text features. IPSJ SIG Notes (94-FI-33). 1994;94:17-22 (in Japanese, with English abstract)

29. Schamber, L. Eisenberg MB, Nilan, MS. A re-examination of relevance : toward a dynamic, situational definition. Inf Proc Manag. 1990;26:755-776

30. Schamber, L. "1. Relevance and information behavior". Annual Review of Information Science and Technology. Vol.29, 1994, p.3-48

31. Wang, P. Users' information needs at different stages of a research project: a cognitive view. Paper presented in ISIC '96. 1996,19p

32. Sackett, DL, Haynes, RB, Tugwell, P. Clinical epidemiology : a basic science for clinical medicine. Boston, Little Browns. ,1985,370p.

33. Saracevic, T. Relevance reconsiderd '96. Information science : integration in perspective. Proceedings CoLIS2, 1996, p.201-218

34. Ingwersen, P. Cognitive perspectives of information retrieval interaction : elements of a cognitive IR theory. J Doc. 1996; 52: 3-50

35. Salton, G, Allan, J, Buckley, C. Approaches to passage retrieval in full text information systems. Proceeding of 16th ACM-SIGIR. 1994, p.49-58.

36. Swanson, DR. Undiscovered public knowledge. Library Quarterly. 1986;56:103-118

37. Rau, LF, Jabobs, P. SCISOR : Extracting information from on-line news. Comm ACM. 1990;33:.88-97

38. McKeown, K, Radev, DR. Generating summaries of multiple news articles. Proceeding of 18th ACM/SIG-IR, 1995, p.74-83

39. Kurata, K. (1991) The Relation between information media generation process and informational content. Paper presented in Annual Meeting of Mita Society for Library and Information Science. (in Japanese)

40. Suga, A. (1992) Information retrieval based on the logical information in the documents ; application of the Categories for information retrieval. Unpublished Graduation Thesis, Keio University, Tokyo, Japan. (in Japanese)

41. Kawaharamura, M. (1992) The validity and application of structure analysis of information media using the Categories : towards enhanced utilization of the primary information . Unpublished Graduation Thesis, Keio University, Tokyo, Japan, 69p. (in Japanese)

42. Day, RA. How to write and publish a scientific paper, 3rd ed. Oryx Press, 1988, 211p.

43. Gonnet, GH, Baeza-Yates, RA, Snider, T. “5. New indices for text : PAT trees and PAT arrays”. Information retrieval : data structures & algorithms. Edited by Frakes, WB, and Baeza-Yates, R. Englewood Cliffs, N.J., Prentice Hall, 1992, p.66-82.

44. Livelink index engine query language reference. OpenText Corporation, 1996.

Appendix 1 : Comparison of Passages Extracted from Texts (English translation)

(a) the difference in the effect of interferon (IFN) therapy with age

English translation of extracted passage from #1

Secondly, we studied factors in 10 patients whose s-GPT were normalized after Interferon treatment and in 4 whose s-GPT were not normalized among the 14 patients (10 with continuous administration and 4 with intermittent administration)(Table 1). No significant difference was found between the two groups in terms of sex, age, history of blood transfusion, general liver function tests or serum OD level for anti-HCV.

English translation of extracted passage from #2

Multivariate analysis of the factors influencing the normalization of aminotransferase with IFN therapy was conducted. Table 5 presents the results of the studies on the parameters of age, sex pretreatment histology, administration of IFN (continuous administration for 4 weeks, 6 weeks or more, continuous and intermittent, intermittent). Significant differences were found in 5 of these factors. These were the total dosage of IFN, age when the treatment started, administration of IFN, and pretreatment histology. The normalization ratio of GPT by IFN was higher in the patients whose total IFN dosage was more than 400MU, whose age was less than 35, with continuous and intermittent administration, whose pretreatment histology was CH2A, and who were female.

English translation of extracted passage from #9

Next, we studied the patients who responded to IFN treatment and those who did not(Table 5). For the analysis of relevant factors, no significant difference was found with regard to age, sex, or pretreatment level of GPT but the number of patients with CAH2B was significantly larger among those who did not respond than those who did.

English translation of extracted passage from #27

3. Analysis of the patient factors in those who responded after 6 month-intermittent administration of IFN and those who did not(Table 1)

The age of those who responded was significantly younger ($p<0.01$), 40.3±12.3 years versus 50.1±7.7 years for the nonresponders. The periods between blood transfusion and IFN treatment, and the duration of hepatitis were significantly shorter for the responders ($p<0.05$), 66±112.3 months and 24±33.2 months, respectively, versus 218.9±152.4 and 62.8±37 months for the nonresponders

English translation of extracted passage from #43

1. Univariate analysis between the groups(table 2)

The ages of the responders ranged from 22 to 28, 42±14.5 years old on average, *i.e.* significantly younger than the nonresponders who were from 51 to 67, 58.1±5.7 years old on average($p<0.01$).

Among the patients who were anti-HCV positive, significant differences were found between those who responded to the treatment and those who did not in terms of age, pretreatment 2-5 AS activity, maximum 2-5AS activity, and serum GPT level.

We conducted multivariate discriminate analysis using 5 variables including age, pretreatment 2-5 AS activity, maximum 2-5 AS activity, and the increment ratio of 2-5 AS. Significant differences were found between the responders and nonresponders in the latter., Furthermore, the serum GPT level tended to be lower in the responders. The discriminate function was; y = - 2.51 x 10^{-1} (age) + 6.7 x 10^{-3} (pretreatment 2-5 AS activity) - 2.1 x 10^{-2} (2-5 AS activity) + 6.2 x 10^{-2} (increment ratio of 2-5 AS) - 2.7 x 10^{-2} (GPT) + 1.583.

(b) anti-HCV positive rate

English translation of extracted passage from #10

It has been reported that the anti-HCV positive rate in Japanese is 1.1% to 1.2% based on surveys of blood donors[KATA90a].

English translation of extracted passage from #18
It has also been reported that the ratio of anti-HCV carriers among blood donors in Japan is about 1%, which is almost the same or slightly less than the 1% of anti-HBV carriers and is consistent with the ratio of anti-HCV carriers in blood donors in Yamagata prefecture being 1%[TSUC90].
English translation of extracted passage from #19
The anti-HCV positive rate in blood donors at the Japan Red Cross Blood Center has been reported [NISH89]. That of chronic dialysis patients is markedly elevated in comparison. As shown in Table 4, the anti-HCV positive rate in the U.S. and other European countries ranges from 0.2% to 1.4% while that one in Germany is slightly lower than the others.
English translation of extracted passage from #33
The anti-HCV positive rate in patients with a history of blood transfusion was as high as 28% and even among those with no such history, the rate was 6.7%, markedly higher than the rate of ordinary blood donors[SHIK91].
English translation of extracted passage from #38
It has been reported that the anti-HCV positive rate in blood donors is 1-2 % and that viral hepatitis C is caused by the blood transfusion of anti-HCV positive blood[KHOO89][KATA90b].

The results in the original language (Japanese) is shown in Appendix 1'.

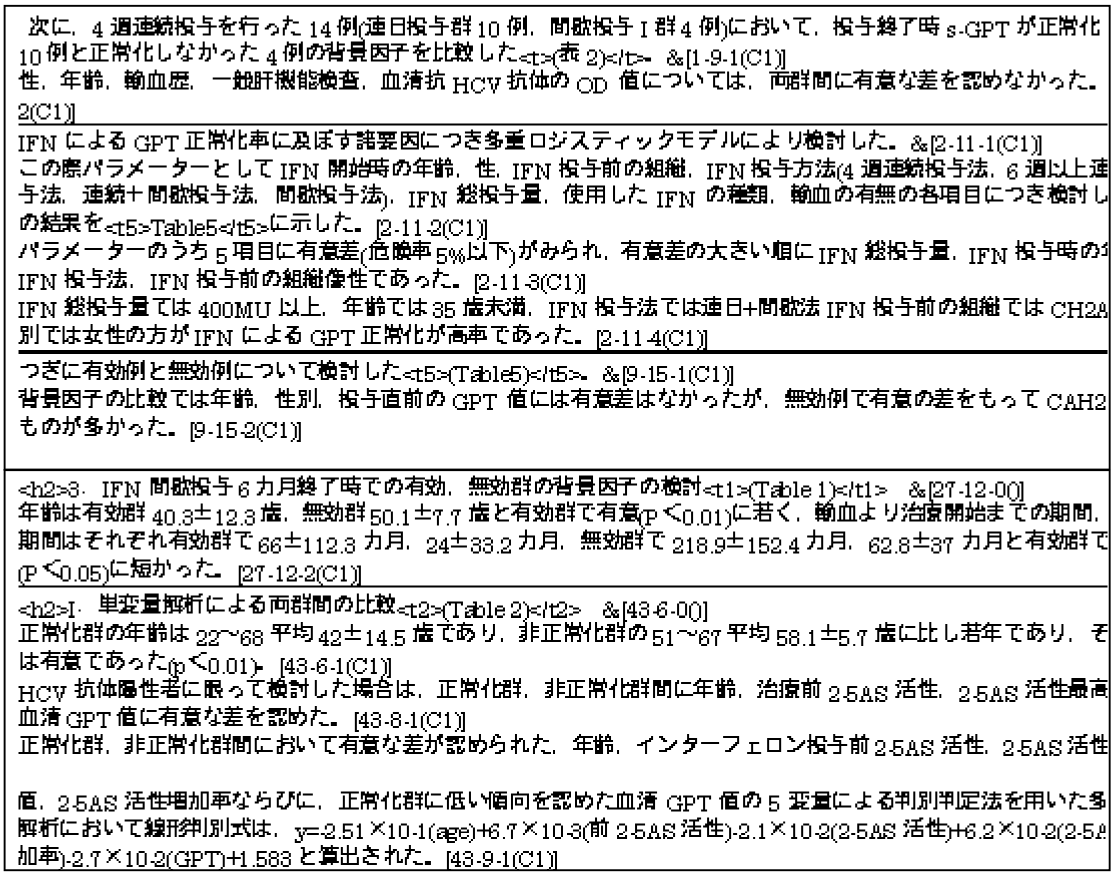
Appendix 2 : Examples of Extracted "State-of-Art" Sentences (English translation)

[English Translation of Extracted Passages from #9]
It has been revealed that the majority of cases diagnosed as non A non B hepatitis are anti-HCV positive by using the hepatitis C virus antibody (hereafter, anti-HCV antibody) which was developed by Ortho, Co. based on Chiron Co.'s recombinant C-100 clone. However, this antibody assay kit has certain problems, such as, (1) its reproducibility. (2) data reliability when the OD level is higher than 2.0 with the EIA method. Thus, the correlation with the anti-HCV level can not be examined when the OD level is higher than 2.0.
[English Translation of Extracted Passages from #17]
Since discovery of the hepatitis C virus, the diagnosis of hepatitis C can be made based on C-100 antibody [CHOO89] [KUHG89], and the findings of hepatitis C have thus increased remarkably [MIYA89] [ESTE89] [POEL89]. Furthermore, hepatitis C virus RNA (HCV RNA) can be detected by the polymerase chain reaction (PCR) [WEIN90] [HAYA90] [KANE90].
[English Translation of Extracted Passages from #26]
Although it has been said that more than 90% of cases with post blood transfusion hepatitis are caused by non A non B viral hepatitis, its nature was only recently elucidated. A research group at Chiron, Co., in the U.S., analyzed cDNA separated from the serum of chimpanzees infected with NANB virus, using a genetic engineering technique and revealed it to be from the genome of the hepatitis C virus (HCV). Furthermore, Kuo et al. [KHOG89] developed an assay system based on the protein which was prepared by fermenting a part of the cDNA.. For acute non A non B hepatitis, it has been demonstrated that anti-HCV is not detectable in the very early stage of the disease but rather can be detected after 3-6 months. *i.e.* this assay was found to be inadequate for diagnosis at the beginning of the disease course of acute hepatitis.
[English Translation of Extracted Passages from #8]
Since Prince et al.[PRIN74] proposed the existence of non A non B type hepatitis after blood transfusion, etiologic and clinical aspects of the disease have been studied. However the investigations of the virus have not yet succeeded. Recently, an essential investigation was successfully carried out by a research group at Chiron, Co., in the U.S. Choo et al.[CHOO89] analyzed cDNA which had been separated from serum of the chimpanzees infected with NANB virus, using a genetic engineering technique, revealing the DNA to be from the genome of the hepatitis NANB virus (hepatitis C virus, HCV). Furthermore, Kuo et al.[KHOG89] detected viral antibodies fin serum of patients using C100-3, which was prepared utilizing fermented cDNA with yeast.

The results in the original language (Japanese) is shown in Appendix 2’.

Appendix 1’: Comparison of Passages Extracted from Texts (Original)

(a) the difference in the effect of interferon (IFN) therapy with age (in Japanese)



次に，4 週連続投与を行った 14 例(連日投与群 10 例，間歇投与 I 群 4 例)において，投与終了時 s-GPT が正常化
10 例と正常化しなかった 4 例の背景因子を比較した<t>(表 2)</t>。 &[1-9-1(C1)]
性，年齢，輸血歴，一般肝機能検査，血清抗 HCV 抗体の OD 値については，両群間に有意な差を認めなかった。
2(C1)]

IFN による GPT 正常化率に及ぼす諸要因につき多重ロジスティックモデルにより検討した。 &[2-11-1(C1)]
この際パラメーターとして IFN 開始時の年齢，性，IFN 投与前の組織，IFN 投与方法(4 週連続投与法，6 週以上連
与法，連続+間歇投与法，間歇投与法)，IFN 総投与量，使用した IFN の種類，輸血の有無の各項目につき検討し
の結果を<t5>Table5</t5>に示した。 [2-11-2(C1)]
パラメーターのうち 5 項目に有意差(危険率 5%以下)がみられ，有意差の大きい順に IFN 総投与量，IFN 投与時の
IFN 投与法，IFN 投与前の組織像性であった。 [2-11-3(C1)]
IFN 総投与量では 400MU 以上，年齢では 35 歳未満，IFN 投与法では連日+間歇法 IFN 投与前の組織では CH2A
別では女性の方が IFN による GPT 正常化が高率であった。 [2-11-4(C1)]

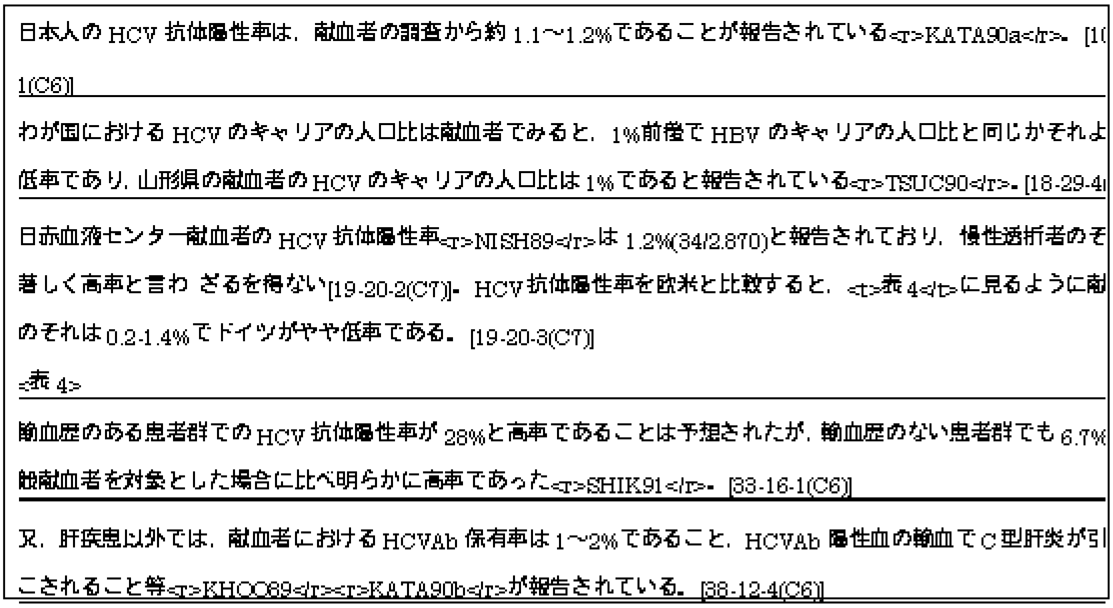
つぎに有効例と無効例について検討した<t5>(Table5)</t5>。 &[9-15-1(C1)]
背景因子の比較では年齢，性別，投与直前の GPT 値には有意差はなかったが，無効例で有意の差をもって CAH2
ものが多かった。 [9-15-2(C1)]

<h2>3．IFN 間歇投与 6 カ月終了時での有効，無効群の背景因子の検討<t1>(Table 1)</t1> &[27-12-0()]
年齢は有効群 40.3±12.3 歳，無効群 50.1±7.7 歳と有効群で有意(P <0.01)に若く，輸血より治療開始までの期間，
期間はそれぞれ有効群で 66±112.3 カ月，24±33.2 カ月，無効群で 218.9±152.4 カ月，62.8±37 カ月と有効群で
(P <0.05)に短かった。 [27-12-2(C1)]

<h2>I．単変量解析による両群間の比較<t2>(Table 2)</t2> &[43-6-0()]
正常化群の年齢は 22~68 平均 42±14.5 歳であり，非正常化群の 51~67 平均 58.1±5.7 歳に比し若年であり，そ
は有意であった(p <0.01)。 [43-6-1(C1)]
HCV 抗体陽性者に限って検討した場合は，正常化群，非正常化群間に年齢，治療前 2-5AS 活性，2-5AS 活性最高
血清 GPT 値に有意な差を認めた。 [43-8-1(C1)]
正常化群，非正常化群間において有意な差が認められた，年齢，インターフェロン投与前 2-5AS 活性，2-5AS 活性

値，2-5AS 活性増加率ならびに，正常化群に低い傾向を認めた血清 GPT 値の 5 変量による判別判定法を用いた多
解析において線形判別式は，y=-2.51×10-1(age)+6.7×10-3(前 2-5AS 活性)-2.1×10-2(2-5AS 活性)+6.2×10-2(2-5A
加率)-2.7×10-2(GPT)+1.583 と算出された。 [43-9-1(C1)]

(b) anti-HCV positive rate (in Japanese)



日本人の HCV 抗体陽性率は，献血者の調査から約 1.1~1.2%であることが報告されている<T>KATA90a</T>。 [1(
1(C6)]

わが国における HCV のキャリアの人口比は献血者でみると，1%前後で HBV のキャリアの人口比と同じかそれよ
低率であり，山形県の献血者の HCV のキャリアの人口比は 1%であると報告されている<T>TSUC90</T>。[18-29-4

日赤血液センター献血者の HCV 抗体陽性率<T>NISH89</T>は 1.2%(34/2.870)と報告されており，慢性透析者のそ
著しく高率と言わ ざるを得ない[19-20-2(C7)]。 HCV 抗体陽性率を欧米と比較すると，<t>表 4</t>に見るように献
のそれは 0.2-1.4%でドイツがやや低率である。 [19-20-3(C7)]
<表 4>

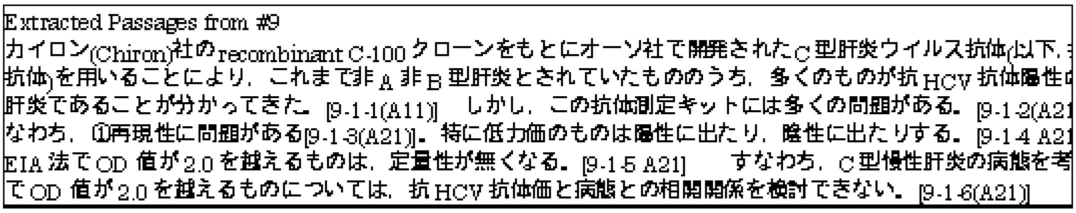
輸血歴のある患者群での HCV 抗体陽性率が 28%と高率であることは予想されたが，輸血歴のない患者群でも 6.7%
般献血者を対象とした場合に比べ明らかに高率であった<T>SHIK91</T>。 [33-16-1(C6)]

又，肝疾患以外では，献血者における HCVAb 保有率は 1~2%であること，HCVAb 陽性血の輸血で C 型肝炎が引
こされること等<T>KHOC89</T><T>KATA90b</T>が報告されている。 [38-12-4(C6)]

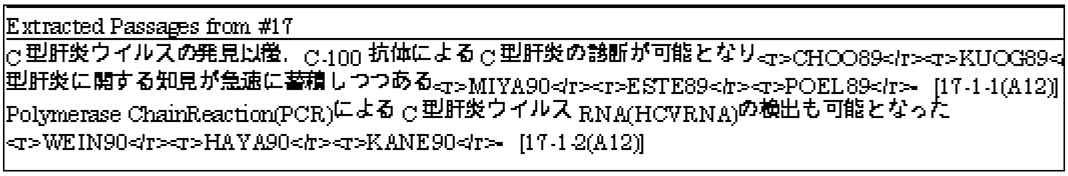
In [dd-pp-ss(cat)], *dd* represents the document number, *pp* represents the paragraph number in a document, *ss* represents the sentence number in the paragraph, and *cat* represents the Categories assigned to the sentence.

Appendix 2’: Examples of Extracted “State-of-Art” Sentences (Original)

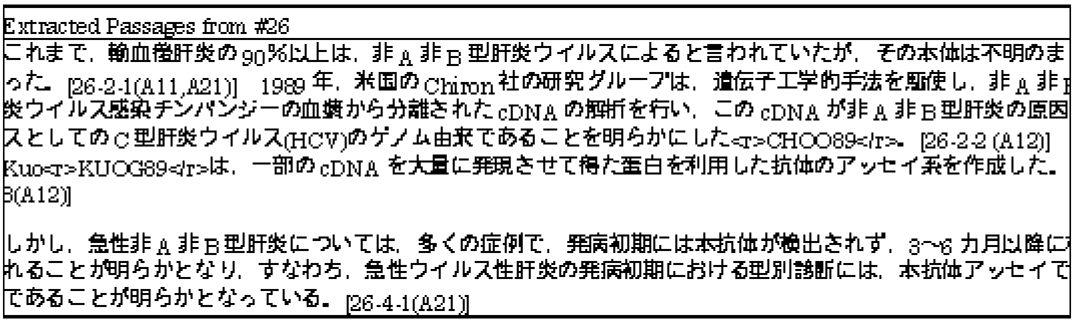
(in Japanese)



Extracted Passages from #9
カイロン(Chiron)社のrecombinant C-100クローンをもとにオーソ社で開発されたC型肝炎ウイルス抗体(以下，
抗体)を用いることにより，これまで非A非B型肝炎とされていたもののうち，多くのものが抗HCV抗体陽性
肝炎であることが分かってきた。 [9-1-1(A11)]　しかし，この抗体測定キットには多くの問題がある。 [9-1-2(A21
なわち，①再現性に問題がある[9-1-3(A21)]。特に低力価のものは陽性に出たり，陰性に出たりする。 [9-1-4 A21
EIA法でOD値が2.0を越えるものは，定量性が無くなる。 [9-1-5 A21]　すなわち，C型慢性肝炎の病態を考
てOD値が2.0を越えるものについては，抗HCV抗体価と病態との相関関係を検討できない。 [9-1-6(A21)]



Extracted Passages from #17
C型肝炎ウイルスの発見以後，C-100抗体によるC型肝炎の診断が可能となり<r>CHOO89</r><r>KUOG89<
型肝炎に関する知見が急速に蓄積しつつある<r>MIYA90</r><r>ESTE89</r><r>POEL89</r>。 [17-1-1(A12)]
Polymerase ChainReaction(PCR)によるC型肝炎ウイルスRNA(HCVRNA)の検出も可能となった
<r>WEIN90</r><r>HAYA90</r><r>KANE90</r>。 [17-1-2(A12)]



Extracted Passages from #26
これまで，輸血後肝炎の90%以上は，非A非B型肝炎ウイルスによると言われていたが，その本体は不明のま
った。 [26-2-1(A11,A21)]　1989年，米国のChiron社の研究グループは，遺伝子工学的手法を駆使し，非A非
炎ウイルス感染チンパンジーの血漿から分離されたcDNAの解析を行い，このcDNAが非A非B型肝炎の原因
スとしてのC型肝炎ウイルス(HCV)のゲノム由来であることを明らかにした<r>CHOO89</r>。 [26-2-2 (A12)]
Kuo<r>KUOG89</r>は，一部のcDNAを大量に発現させて得た蛋白を利用した抗体のアッセイ系を作成した。
3(A12)]

しかし，急性非A非B型肝炎については，多くの症例で，発病初期には本抗体が検出されず，3~6カ月以降に
れることが明らかとなり，すなわち，急性ウイルス性肝炎の発病初期における型別診断には，本抗体アッセイで
であることが明らかとなっている。 [26-4-1(A21)]

In [dd-pp-ss(cat)], *dd* represents the document number, *pp* represents the paragraph number in a document, *ss* represents the sentence number in the paragraph, and *cat* represents the Categories assigned to the sentence.