

コンコーダンスを指向したテキストデータベースの研究

坂口基彦、早川栄一、並木美太郎、高橋延匡

東京農工大学工学部電子情報工学科

本稿では、コンコーダンスを指向したテキストデータベースの設計と実現について述べる。従来文学、歴史学などの研究者が紙面上のコンコーダンスを用いておこなっていた文書の参照、コンコーダンスの作成の過程を指向して、計算機上のテキストデータに対して、コンコーダンスを作成、コンコーダンスを用いた内容の参照などの一連の作業を通して行えるテキストデータベースを設計、実現した。このシステムは(1)従来時間のかかったコンコーダンスの作成の手間を文字列検索などの計算機の支援によって軽減する、(2)コンコーダンスへの書込みなど個人的情報が管理できる、などの特徴を備える。また本システムを利用して夏目漱石の『坊っちゃん』、琉球王国の外交文書『歴代宝案』の読下し文の一部、についてコンコーダンスを作成し、内容を参照した結果について報告する。

Study of a Text-Database Pointing to a Concordance

Motohiko SAKAGUCHI, Eiichi HAYAKAWA, Mitarou NAMIKI
and Nobumasa TAKAHASHI

Department of Computer Science, Faculty of Technology,
Tokyo University of Agriculture and Technology

This paper describes the design and implementation of a text-database pointing to a concordance. This database is designed to point to text-reference works of scholars of literature and history with a concordance on paper. Using this database, users can make concordances and refer to text-data with concordances. This database has two characteristics:(1)it can reduce the cost of making a concordance with the support of a computer, such as string searches, (2)users can accumulate knowledge of text into a concordance by writing notes.

We also report the result of making concordances and references to text about "Bocchan" and a diplomatic document, "Rekidai-Hoan" from Ryukyu, with this database.

1. はじめに

従来、紙面上で扱われていた文書が計算機上のテキストデータとして扱われることが多くなった。テキストはデータベースのデータのように検索することを考慮して書かれておらず、必要な内容の記述箇所を探すのに手間がかかる。そこでテキストから必要な内容を効率よく参照することへの要求は高い。また効率よく参照するためには、どのように参照されるかを考慮してテキストに含まれる情報を整理しておく必要がある。

文学、歴史学などの研究者は、研究対象の文献のコンコーダンス（用例索引）を、参照したい内容の記述箇所を探すために利用し、研究に役立てている。コンコーダンスは、語句の出現する位置と文脈が記載されている索引である。研究者はコンコーダンスに調べた結果を書込み、内容の参照に役立てている。しかしコンコーダンス自体が大きく、効率よく必要な項目を探せない、参照したい文書にコンコーダンスがない場合には、文脈の切り出しなど作成に非常に時間がかかるなど問題点も多い。

従来この種の研究は、KWIC の作成に重点をおくことが多い。目的が従来の紙面上のコンコーダンスを計算機の補助を用いて作成することであったためと考えられる。そのためコンコーダンスから必要な項目を検索し、テキストの実体を参照することや、コンコーダンスで重要な情報の書込みを支援できない。そこで著者らは、コンコーダンスの作成、コンコーダンスを用いた内容の参照、結果の書込みなどの実際のコンコーダンスを用いた一連の作業を通して計算機で補助するテキストデータベースシステムを考案した。

本稿では、コンコーダンスを指向したテキストデータベースの設計と実現について述べる。またシステムを用いて、夏目漱石の『坊っちゃん』[4]、琉球外交文書『歴代宝案』の第1集抄の中の100通の文書の読下し文に対してコンコーダンスを作成し、内容の参照を行なった結果について述べる。

2. システムの概要

2.1 従来のコンコーダンスの問題点

従来のコンコーダンスを用いた作業の問題点として次のことが上げられる。

(1) 必要なコンコーダンスが存在しない

文献を研究するときに、コンコーダンスが存在しないことが多い。逆に言えば、すでにコンコーダンスが作成されている文献は研究がある程度終わっている。そのためコンコーダンスを作成する必要性は高い。しかし文脈の切り出しなど大変時間がかかる作業である。また存在しても使用者の要求にあわなければ、役に立たない。そのため使用者のニーズにあったコンコーダンスを簡単に作成できる必要が

ある。

(2) 文献を引くなどの作業に時間がかかる

膨大な項目の中から必要な項目を探したり、参照する箇所を文献から引く作業に時間がかかる。この時間は文献の参照という目的に対して、よけいな時間であり、参照の効率を落とす原因となっている。

2.2 システムの特徴

本システムは、次の特徴を持つ。

(1) コンコーダンスの参照、作成過程を反映できる

従来のテキストデータベースでは、テキスト中に参照したい語句があった場合、検索システムにその語句をキーとして入力する。そして結果の KWIC で参照位置を調べ、テキストの参照箇所を引く必要があった。この形態には無駄が多く、従来のコンコーダンスを使用する過程を反映できない。また作成に関しても、テキスト中に重要な語句があった場合、その場で項目を作成できれば、より使用者のニーズをコンコーダンスに反映できる。

本システムでは、参照作業で一番重要であるテキストを読む作業を重視し、テキストビューアと検索システム、コンコーダンス作成システムを融合し、従来のコンコーダンスを用いた作業を反映させる。具体的には、テキスト中の語句を指定すれば自動的に項目を表示する。また参照したい箇所の KWIC を選択することで、自動的にテキストのその部分を表示する。テキスト中の重要な語句を指定すれば、自動的に文字列検索を行いその語句の出現箇所を登録、KWIC を作成し項目として登録するといった機能を持つ。

(2) 個人の要求に合うコンコーダンスを作成できる

コンコーダンスは作成に時間がかかり、手軽に個人レベルで作成することは難しかった。特に時間がかかるのは、語句の出現位置の探索と、文脈の切り出しであった。本システムでは、計算機の文字列検索の支援により出現位置探索、文脈切り出しを自動化する。作成者は、項目にする語句の切り出しの作業のみを行えばよいので、ニーズにあったコンコーダンスを作成することができる。

また作成者は、そのテキストに様々な知識を持つ。知識を頭の中の知識としてだけでなく、記述しておくことは、コンコーダンスの重要な役目であった。そのためコンコーダンスの項目に語句の意味、調べて分かった点など情報を書き込むことができる。

2.3 設計方針

システムの設計方針を次のように定める。

(1) 文書に対するユーザの知識を付加できる

文書は記述上には現れない様々な情報がある。例えば語句の意味などである。その他にも語句間のつながりは、紙の上で表現することは難しいが、計算

機ではリンクを張って表現することができる。またコンコーダンスを用いて参照した結果、分かったことなどを書き込めれば、研究に役立つ。

(2) データとしてブレンテキストを扱う

テキストのデータベースは大きく二つあり、一つは著者、文書名、キーワードから必要な文献を探すもの、もう一つは、加工していないテキストを検索するものである。文学、歴史学などでは、文献の中の具体的記述が重要である。そのため本システムは後者を扱う。テキストは様々な物理構造、論理構造を持つが、本システムでは最も基本的な形式であるブレンテキストを対象とする。とくに文学、歴史学などの分野では著作権の問題上ブレンテキストがもっとも適当である。

(3) テキストとコンコーダンスのデータを分離する

大きいテキストはCD-ROMなどの書込みのできないメディアで配布されることが考えられる。そのためテキストデータ自体には、コンコーダンスに関するデータの追加、書込みなどはおこなわず、分離したデータとして扱う。

(4) コンコーダンス作成の機械的作業を自動化する

目的でも述べたがコンコーダンス作成に一番時間がかかるのは、語句の出現位置の探索と、その文脈の切り出しである。これらの作業を計算機で支援する。

3. システムの設計

3.1 システムの全体構成

システムの全体構成を図1に示す。

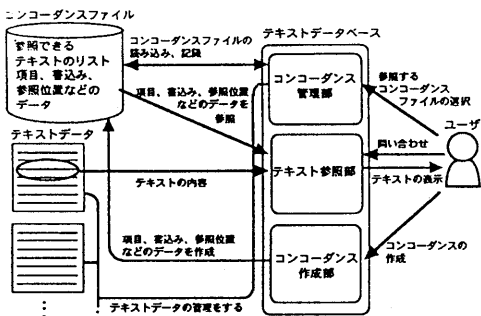


図1 システムの全体構成

データは、次の二つからなる。

(1) コンコーダンスファイル 従来の紙面上のコンコーダンスに当たるデータである。中身はそのコンコーダンスで参照できるテキストのリスト、コンコーダンスの項目、参照位置、書込みなどが含まれている。KWICは含まず、記録されている参照位置が

ら参照時に作成する。

(2) テキストデータ 内容を参照するブレンテキストである。コンコーダンスによって内容を参照できるテキストデータをあらかじめ登録しておく。

本システムの特徴は、コンコーダンス作成からコンコーダンスを用いた参照など一連の作業を通して支援することである。その流れに沿って各部について説明する。

(1) 参照するテキストを管理する。

コンコーダンスは一文献に一つ作成されるのではなく、一つの分野に関して複数の文献から作成される。コンコーダンスで参照するテキストをシステムで管理しておく。コンコーダンスファイルに参照するテキストのファイル名を格納しておき、ファイルを読み込んだときに自動的にシステムに登録する。

これをコンコーダンス管理部でおこなう。

(2) コンコーダンスを作成する。

本システムのコンコーダンスの作成は、まずテキストビューアを用いてテキストの中からコンコーダンスの項目として作成する語句を探す。次に項目の参照位置を文字列検索を用いて自動登録する。例えば、歴史関係の文献では、同一人物に名前、役名、改名したなどで様々な表記をされる場合がある。同一人物であるので一つの項目として参照できた方がよい。そのため項目を作成するときに、項目名とその参照位置検索用の検索条件を別に入力できる。検索条件はAND,OR検索に対応する。

参照位置の登録は三通りの方法でおこなう。(i) 指定された検索条件で文字列検索をおこない自動的に参照位置を登録する(自動作成)。(ii) 単純に文字列検索をおこなうと、必要ない部分まで参照位置として登録される。例えば「自動」で「自動車」が検索される。あらかじめ誤った文字列検索が多くおこなわれそうな場合は、検索された箇所を表示して参照位置として登録するか作成者に判断させる(半自動作成)。(iii) 項目名をラベルとして扱い、参照位置を作成者が手動で登録する(手動作成)。例えばテキスト中で非常に大事な箇所を参照するために、「重要点」という項目を作り重要な箇所を作成者がテキストの中から指定する。

また参照位置登録後に項目に対して関連のある項目を関連項目として登録し、項目に関する知識を書き込むことができる。

これをコンコーダンス作成部で行う。

(3) コンコーダンスを用いて参照する。

参照する項目の検索方法として、従来のキーを入力して検索する方法以外に、表示されているテキスト中の語句を指定する、項目がメニュー形式で表示されているものから選択するといった従来の作業を反映した方法で検索を行える。項目を選択すると

KWIC や書込み、関連項目が表示され、KWIC をもとに参照したい箇所を特定する。そして参照したい箇所の KWIC を選択することで、自動的にそのテキストの実体が表示される。

これをテキスト参照部で行う。

3.2 コンコーダンスの設計

項目の例として夏目漱石『坊っちゃん』の「赤シャツ」の項目を図2に示し、各部分について説明する。

『坊っちゃん』コンコーダンスの「赤シャツ」

(1) 項目名

(2) 検索条件

第1候補	<input type="text" value="赤シャツ"/>	と	<input type="text"/>
第2候補	<input type="text" value="教頭"/>	と	<input type="text"/>
第3候補	<input type="text"/>	と	<input type="text"/>

(3) 書込み

坊っちゃんが勤める中学の教頭。いつも赤シャツを着ていることから坊っちゃんがこのあだ名を付けた

(4) 関連項目

関連項目	関連内容
文学士	赤シャツは文学士である
琥珀色のパイプ	赤シャツが愛用している

(5) 参照位置

テキスト名	構造名	行数	KWIC
坊っちゃん	3章	74行	いたから、誰も赤シャツもちっとも
坊っちゃん	4章	46行	固るな、校長が教頭に出来うとめん
坊っちゃん	5章	31行	行きませんかと赤シャツがおれに講
坊っちゃん	5章	31行	えて、停留所で赤シャツと野だにい

図2 コンコーダンスの項目の例

(1) **項目名** 項目の名前である。従来のコンコーダンスでは、項目はテキスト中に出現する語句である。しかしテキストの内容を自由に参照するためにはこれでは制限がある。そのため項目名をラベルとして扱い、(2) **検索条件** と分けて定義する。

(2) **検索条件** 項目の参照位置を文字列検索を用いて登録する条件である。候補は、OR 検索に対応するものである。また「と」はAND 検索に対応し、複数の語句が連続してあられる場所を検索する。

(3) **書込み** 項目に対する書込みである。例えば項目への説明や、コンコーダンスを用いて参照し分かった点などを自由に書き込む。

(4) **関連項目** 項目に関連する項目である。辞書でいう参照に当たるものである。リンクを張るだけでは、項目間にどんな関連があるかという重要な情報が抜ける。そのため関連内容と言う情報で表す。

(5) **参照位置** その項目で参照するテキスト中の位置である。参照の補助として次の情報を持つ。

- テキスト名 参照位置があるテキスト名
- 構造 参照位置の章、節などの論理構造
- 行数 参照位置が存在する行数
- KWIC 参照位置をKWIC形式で切り出したもの

4 各部の機能

4.1 コンコーダンス管理部

主な機能をあげる。

(1) コンコーダンスファイルの読み込み、記録

コンコーダンスファイルを読み込むと、コンコーダンスのデータが読み込まれ、そのコンコーダンスで参照できるテキストが自動的に登録される。また作成したコンコーダンスを記録する。

(2) 参照テキストの登録

コンコーダンスの対象となるテキストをあらかじめ登録しておく。登録されているテキストデータがコンコーダンス作成のとき検索対象となる。一つのコンコーダンスに複数のテキストを登録しておくことで従来は面倒であった複数のテキストからの参照を効率よくおこなう。

4.2 テキスト参照部

テキスト参照部では、コンコーダンスを用いてテキストの参照をおこなう。参照する項目を、実際の参照作業を考えた次の三通りで選択する。

- 表示されているテキストの語句を指定
- 参照したい項目をキーボードで入力
- メニュー形式で表示された項目名一覧から選択

本システムで項目は、項目名と検索条件を別のものと定義している。そのため(1)(2)の方法で指定された場合は、図2の例では、「赤シャツ」「教頭」のどちらの語句が指定されても「赤シャツ」の項目を参照できる。

そしてこれらの方法で指定された項目のKWIC、書込み、関連項目を表示し、最終的に参照するテキストの箇所を選択する。選択した箇所が必要としない箇所であることが多いと考えられるので、容易に箇所の選択と表示を繰り返せる。

テキスト参照の流れを図3に示す。

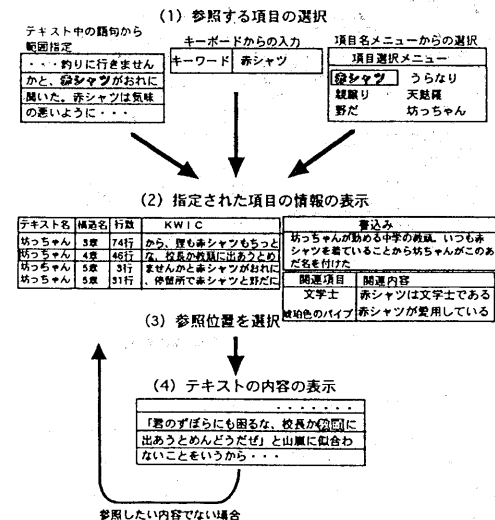


図3 テキスト参照の流れ

4.3 コンコーダンス作成部

主な機能をあげる。

(1) 項目作成

本システムでは、項目名と検索条件を別々に定義し、項目名はただのラベルである。項目の参照位置の登録方法を次の三つから選択する。

(i) 自動作成 検索条件から自動的に登録する（主に項目名=検索条件の時）

(ii) 半自動作成 検索した結果を作成者に提示し参照位置として登録するかを判定させる（検索に必要としないものが高確率で含まれる可能性が高い場合）

(iii) 手動作成 検索で自動的に参照位置を登録できない場合、検索条件は可能性の高い場所の検出に用いて、最終的に作成者の手で参照位置を登録する

(3.2 (2) の「重要点」のような項目)

図4に示す、項目作成画面に作成型、項目名、検索条件を入力すると指定された型によって参照位置の登録をおこなう。

図4 項目作成画面

(2) 項目自動作成

テキスト中の語句を範囲指定し、その語句を項目名=検索条件として項目を自動作成する。本システムでは、項目にする語句をテキスト中から作成者が探す。この方法で頻りに項目が作成されるので手間を軽減できる。

(3) 書き込み

項目に対して、語句の意味、コンコーダンスを用いて分かった点などをまとめるなど、自由に書き込みをおこなえる。

(4) 関連項目作成

関連する項目どうしにリンクを張る。またその項目間にどのようなつながりがあるかを関連内容として書き込むことができる。

4.4 その他の機能

(1) 構造の定義

著作権問題などの理由から本システムでは、プレーンテキストを対象としている。しかし章、節などの論理構造は参照時に重要な情報である。例えば語句の出現位置が、本文であるのか注であるのかでは意味が違う。そこで構造をシステムで管理する。構造名(X章など)を入力し、テキスト中の構造の先頭位置を作成者が指定する。その位置から次の構造

の先頭までの範囲がその構造となる。

(2) コンコーダンスの紙面へ出力

本システムは計算機上でコンコーダンスを扱うが、紙面で利用できる方がよい場合もあるので、従来の紙面の形式でコンコーダンスを出力する。

5 システムの実現

システムの初版を、次の環境で実現した。

動作環境：NEC PC-98 シリーズ

OS：MS-DOS

言語：C 言語

ソース行数：約 8000 行

6 システムの応用例

6.1 『坊っちゃん』コンコーダンス

本システムを用いて夏目漱石『坊っちゃん』のコンコーダンスを作成した。その概要を次に示す。

テキストサイズ：約 180kbyte

項目数：152 項目

作成時間：約 6 時間

総参照位置：約 2000

コンコーダンスの作成に約 6 時間かかった。作成されたコンコーダンスの質に時間は大きく関わるが、約 2000 の文脈を切り出す作業だけでも紙面上でおこなえば明らかにこれ以上の時間がかかる。本システムの方針であるコンコーダンスの作成の機械的作業の自動化によりかなり作業時間を短縮できた。

次に参照例を述べる。『坊っちゃん』は小説であるので、当然語句に表記のゆれがある。例えば「田舎者」という語句は、「田舎者」と「田舎もの」の両方の表記がされる。作成時に項目名を「田舎者」とし、検索条件の第1候補を「田舎者」、第2候補を「田舎もの」として作成した。どちらの表記で出てきても「田舎者」の項目を参照することができた。図5に「田舎者」の項目で「田舎もの」を参照した結果を示す。

項目名	検索条件	構造	行数	備考
坊っちゃん	坊っちゃん第2章	構造:未定義	11行	379行
坊第2章	L 44 だろ。-のくせ			
坊第2章	L 47 わない。-はしみ			書き込み
坊第3章	L 16 しんを-に弱み			
坊第3章	L 117 はない。-はこ			

図5 「田舎者」の参照画面

6.2 『歴代宝案』の読下し文コンコーダンス

『歴代宝案』は琉球の外交文書である。今回は第1集の中の100通の読下し文についてコンコーダンスを作成した。この読下し文は図6に示すように文書内容、その文書の納められている巻、読下し文、語句注からなっている。構造としてこれらを作成した。

[九 中山王尚巴志より仁宗の皇太子あて箋文(一四二五年)]
 [巻12-4](1)
 琉球国中山王臣尚巴志、誠懼誠*、稽首頓首して上言すらく、伏して以て、皇天の眷佑もて豊いなる運を弘く開き、大本益々隆にして臣民忻戴す。敬んで惟うに、皇太子殿下は寛仁、育徳、敬謹、心に存し、万世の基を嗣承す。實みて慮照かき運を奉じ、ここを以て貞符協応し、万邦、永く寧ず。臣尚巴志遠く蕃地に居るも、心馳せて遙かに賀す。前星を仰望して數算を千秋に祝す。瞻仰で激切屏營の至に任うるなし。讀みて箋を奉じて賀を恭し、以て聞す。
 洪熙元年閏七月十七日 琉球国中山王臣尚巴志、謹みて上箋す
 [注1]皇太子 ここでは仁宗の長子の蘭基、後の宜宗宣徳帝のこと。
 2貞符協応 天地が調和する、吉兆が現実となるの意。
 3前星 皇太子の異称。
 4數算……祝す 皇太子の寿命の千年の長きを祝う。]

図6 『歴代宝案』の読下し文

『歴代宝案』コンコーダンスの概要を次に示す。

- テキストサイズ: 約 125kbyte
- 項目数: 214 項目
- 作成時間: 約 9 時間
- 総参照位置: 約 2500

参照例として明の年号の「洪熙」の参照結果を図7に示す。洪熙は仁宗洪熙帝が在位1年で死んだため元年しか存在しない。しかしKWICを見ると二年が見つかる。KWICを選択しその箇所のテキストを参照する。この例では、同じ画面に注が表示されているのでその理由が分かる。離れた場所にあっても、参照位置選択の時に構造が「注釈」である部分を参照すれば高い確率で何らかの情報を得ることができる。このようにKWICや構造の情報は参照時に有効である。

ファイル	フォルダ	ビュー	検索	参照	作成	削除	機能	直接参照	画作成
今、開す!									
鑑金段五匹 赤段二十四!									
銀貨三千斤 今轉じたるは二千五百斤正!									
腰刀五柄 指紙扇三十柄!									
大青盤二十箇 小青盤四百箇 小青碗二千箇!									
右、通國に送す!									
洪熙二年九月初十日!									
[注1]参照 めぐりあふ。友好関係を取り結ぶの意か。]↓									
[※本文の記年の洪熙二年という年号はないが、通國送のためか。]↓									
洪熙の参照位置 表示									
開出項目 開出内容									
歴代00018 読下し文 L 8。符文は一元年十									
仁宗 仁宗洪熙帝在位の年号									
歴代00019 読下し文 L 7。執照は一元年十									
意読み									
歴代00022 注釈 L 24の記年の一二年と									
仁宗洪熙帝在位の年号。一四二五の一年									
歴代00024 読下し文 L 7よるに、一年間は									
間。									
表示(RETURN) の旨(SPACE) ウィンドウ移動(SHIFT)+[←][→] メニュー終了(ESC)									

図7 「洪熙」の参照画面

作業をおこなって得た知見を述べる。著者は歴史に対して素人であり、『歴代宝案』及び琉球の知識がまったくなかった。そのため読下し文とはいえ難解な『歴代宝案』のコンコーダンスを作成する作業は、非常に大変であった。しかしながらコンコーダンスの項目を作成するという作業を通して、少しずつ分かったことを書き込む、関係のある項目にリンクを張るなどをするのでかなり内容の理解に役立った。つまり本システムが、研究対象に関する知識をまとめることができることを実感できた。またテキスト中から語句を直接検索できたので、通常使用しない漢字を使った語句も簡単に参照できた。つまり『歴代宝案』のような通常使われない語句が使われるテキストに対して、テキストの直接検索が特に有効であることを確認できた。

7 おわりに

本稿では、コンコーダンスを指向したテキストデータベースの設計と実現について述べた。またシステムを用いて『歴代宝案』の読下し文の一部についてコンコーダンス作成の機械的作業の自動化により短時間で作成し、種々の内容について幅広く参照を行うことができた。

今後の研究課題としては、原文(一次資料、漢文のテキストなど)との対応、コンコーダンスを用いたイメージの参照、項目となる語句の切り出しの自動化などが上げられる。またコンコーダンスはネットワークファイルの整理、プログラムの読解支援など幅広く応用が可能であると考えられる。

謝辞 本研究は、文部省科学研究費課題番号06208102 重点領域「沖縄の歴史情報研究」により行われた。『歴代宝案』のテキストデータを提供していただき、様々な助言をいただいた大阪市立大学柴山守教授ほか関係各位に謝意を表する。

参考文献

- [1] 柴山 守: 漢文文書の分かち書きと辞書生成について, 情報処理学会, 95-CH-27, 1995
- [2] 桶谷猪久夫: 文書データベースの設計と構築, 情報処理学会, 91-CH-8, 1991
- [3] 高倉倉吉: 続おきなわ歴史物語, おきなわ文庫, 1986
- [4] 夏目漱石: 坊っちゃん, 新潮文庫, 1950