電子コンコーダンスシステムの設計と実現

3N－4

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）

１．はじめに

新聞、本などの文書には、多量で多種の情報が、含まれているため、文書内容を参照する機会は多い。しかし文書から必要な内容を、効率よく参照することは難しい。文書内容の参照を頻繁に行う、歴史学研究者などは、研究の対象の文書（古事記など）のコンコーダンスを用いている[1]。

コンコーダンスとは、用例索引のことで文書中で使われている重要なキーワードの全出現箇所が、文脈の一部とともに記載されている。キーワードは、五十音順などに整理されている。図１に例として、聖書のコンコーダンスの一部を示す。コンコーダンスを利用した参照の特徴は、予め参照される内容を考慮して、項目を作成することで、従来のデータベースやフルテキストサーチでできなかった内容の参照が可能になることである。

本報告では、計算機上でコンコーダンスを作成し、それを使った文書内容の参照を行えるテキストデータベースである、電子コンコーダンスシステムを考案し、初版を作成した。

語句	出現位置		文脈
アーメン	コ1	14:16	どうして-といえる
	啓	3:14	-なるもの、忠実で真実な証人
愛	サ2	1:26	あなたの-は、女の-よりも
	マタ	24:12	大半の者の-が、冷えるでしょう
	ヨハ	15:13	これより大きな-を持つ者
	ロマ	8:39	神の-からわたしたちを引き離し
	ロマ	13:10	-は立法を全うする者なのです
	コ1	13:4	-はねたまず、自慢せず
	コ1	13:13	このうち最大のものは-です
	コ1	16:14	すべてのことを-をもって行い

図１　聖書コンコーダンスの一部

２．電子コンコーダンスシステムの設計

２．１　設計方針

（１）参照するテキストの形態を保存する。

内容の参照方法として、辞書のように、必要な内容だけを切り出すことも、考えられる。しかし切り出すことで文書に含まれる構成などの情報が失われるので、テキスト自体に手を加えない。

（２）テキストの内容の実体を見せる。

従来のコンコーダンスでは、キーワードの出現箇所と、文脈の一部しか示されていないため、具体的内容がわからなかった。このシステムでは、キーワードの出現箇所を調べる時に、文書のその箇所を表示しておき、参照の手助けをする。

（３）コンコーダンスの作成を補助する。

コンコーダンスの作成には、手間がかかるので、検索によるキーワードの位置検出など計算機の特性を活かしコンコーダンス作成の手間を軽減する。

２．２　全体構成

図２に、システムの全体構成図を示し、次に各部の簡単な説明を行う。

（１）コンコーダンスファイル

このファイルをシステムに読み込むことで、コンコーダンスを使用して、文書内容を参照することが可能になる。次の二つのデータからなる。

①登録テキスト管理データ

そのコンコーダンスファイルで、参照できるテキストに関する情報が登録されている。

②コンコーダンスデータ

キーワード、参照位置が登録されている。

（２）テキストデータ

コンコーダンスを用いて、内容を参照できるテキストである。コンコーダンスファイルを読み込んだ時に、テキストデータが自動的に登録される。

（３）コンコーダンス管理部

コンコーダンスファイルの読込み、テキストの登録など、コンコーダンスを利用する準備を行う。

（４）コンコーダンス作成部

コンコーダンスの作成、その補助を行う。

（５）コンコーダンス参照部

コンコーダンスから、参照する項目と参照位置を選択すると、参照する部分が表示される。

２．３　本システムでのコンコーダンス

このシステムでは、コンコーダンスの内容を単純なキーワードの出現位置に限らず、キーワードによる文書内容のインデキシングにも対応させる。キーワードが文書内容を抽象化したものになることもあり、検索でコンコーダンスの項目を作成する時に、項目の名前、検索条件を別にした方がよいと考えた。また項目間の関係付けや、書き込みも行えるようにする。図３に項目の例を示す。

Design and Implementation of an Automatic Text-referencing System by Keyword

Motohiko SAKAGUCHI, Eiichi HAYAKAWA, Mitarou NAMIKI and Nobumasa TAKAHASHI
Tokyo University of Agriculture and Technology

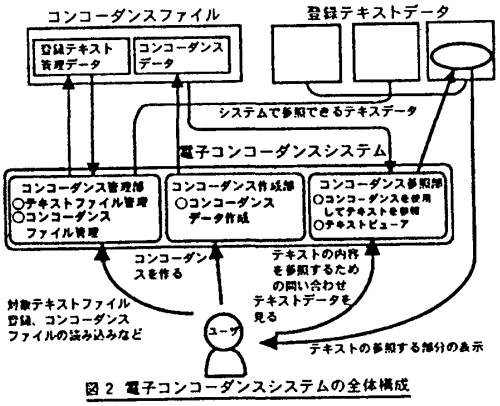

図2 電子コンコーダンスシステムの全体構成

項目		参照位置
項目名	徳川家康	家康第1巻　1頁　5行 家康第5巻　21頁31行 家康第8巻　8頁25行
検索条件	「徳川家康」か「家康」か「行軍」	
関連項目		書き込み
関連項目	関連内容	家康は服従と性格から狸と呼ばれていた。 家康第5巻　21頁31行参照
豊臣秀吉	秀吉の死後天下を取る	
石田三成	関ヶ原の戦いで戦う	

図3 コンコーダンスの構造のイメージ

2．4　本システムの特徴

（1）項目名による参照

例えば、当時の文献から、「関ヶ原の合戦」を参照したい時、「関ヶ原の合戦」という合戦名は、後に付けられたものであり、「関ヶ原の合戦」という用語は使われていない。そのため従来のKWIC では、検索することはできない。しかしコンコーダンスでは「関ヶ原の合戦」の項目に、予め文献に関して知識のある者が、参照位置を登録してあるので、項目名から参照できる。

（2）コンコーダンス作成を補助する。

従来は、人間の手で行われてきたキーワード使用位置の特定を、検索によって自動的に行える。

また上の「関ヶ原の合戦」の項目を作成する時、文献中の「関ヶ原の合戦」に関する場所に、「関ヶ原」、「徳川家康」、「石田三成」などのキーワードが使われていることが考えられる。このキーワードを組み合わせて検索条件とすることで、「関ヶ原の合戦」に関して、記述されている可能性のある部分を検出することができる。作成者は、可能性のある部分から、実際に関係ある部分を選択するだけで、登録できる。

3．電子コンコーダンスシステムの機能設計

3．1　コンコーダンスを用いた文書内容の参照

コンコーダンスから項目を選択し、その参照位置の文書を表示する。項目選択の方法は、項目をメニュー形式で選ぶ方法と、直接キーワードを入力する方法の二通りで行う。メニュー選択では、従来の KWIC で得られない、文書内容も、項目として登録されているので参照することができる。

また項目作成時の検索に使ったキーワードは、重要な情報である。そのため直接入力では、入力されたキーワードが該当する項目だけでなく、検索条件のキーワードが該当する項目も参照候補として挙げる。つまり「関ヶ原の合戦」の例では、「石田三成」を入力した時に、「石田三成」と「関ヶ原の合戦」の項目が参照候補となる。

3．2　コンコーダンスの作成

項目の作成には、2種類の方法を考える。第1の方法は、項目名が単純なキーワードの場合は、キーワードを検索条件として指定することで、出現位置を検索し自動的に作成できる。第2の方法は、項目名が「関ヶ原の合戦」の例のように、文書内容のインデキシングである場合、自動的に検索することは難しいため、AND 、OR 検索を用いて、だいたいの位置を検出し、最終的には作成者の手で参照位置を指定する。

4．初版の作成

システムの初版を、PC-98 上で実現した。現在は、合計で文庫本一冊程度の、複数のテキストデータ(200～300 kbyte)に対して、コンコーダンスの作成、参照ができる。またこのシステムで夏目漱石の『坊っちゃん』のコンコーダンスを作成した。図 4 に、このシステムの特徴である項目名と検索条件の異なる項目の例を挙げる。この項目は、検索条件から、位置を自動的に検出して作成できた。また問題点として、項目にするキーワードの切り出しに手間がかかることが挙げられた。

項目		参照位置
項目名	清への手紙	坊ちゃん第2章　146行
検索条件	「清」か「手紙」	
関連項目		書き込み
関連項目	関連内容	手紙をたくさん書くとあるが、実際に本文中には、この手紙しか記述されていない。
あだ名	あだ名を手紙に書いている	

図4『坊っちゃん』コンコーダンスの項目「清への手紙」

5．おわりに

本稿では、電子コンコーダンスシステムの設計と初版について述べた。文庫本程度の量の文書について、コンコーダンスを作成し、内容を参照することができたので、今後もっと大きな文書（文書の集まり）に対応させ、考察を行う必要がある。

参考文献

[1]星野聰：日本史データベース、情報処理、Vol. 33、No.10、PP.1109-1115、1992