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

4J－2

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）

1. はじめに

文書を計算機上で扱うことが多くなった現在、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。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。

しかし作成時の文脈の切り出しなど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。そのため必要な項目を検索し、テキストの実体を参照することや、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、システムの設計と実現について述べる。また琉球外交文書『歴代宝案』の第１集抄の読下し文にシステムを使用した結果について述べる。

2. システムの概要

2.1 システムの特徴

従来のコンコーダンスを用いた作業の問題点として、(i)作成に時間がかかり使用者のニーズにあったコンコーダンスを作成できない、(ii) 参照する項目を探し、その箇所を引く作業に時間がかかる、などがあった。そのため本システムは、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は、語句を検索システムにキーとして入力し、結果をもとに文書を引いて参照する。この形態では、作業の過程を反映できない。本システムでは、参照作業で一番重要な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、従来の作業を反映させる。具体的には、テキスト中の語句を指定して項目を検索する。KWIC を選択すれば、テキストの実体を表示する、などの機能を持つ。

(2) 個人の要求するコンコーダンスを作成できる

コンコーダンス作成は、語句の出現位置の探索、文脈の切り出しに時間がかかり、個人レベルの作成は難しかった。本システムでは、計算機の支援で出現位置探索、文脈切り出しを自動化する。また作成者のテキストに関する知識を書き込むことができる。

2.2 設計方針

システムの設計方針を次のように定める。

(1) テキストに対するユーザの知識を付加できる

(2) データとしてプレーンテキストを扱う

(3) CD－ROM 等書込み不可のテキストにも対応する

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。

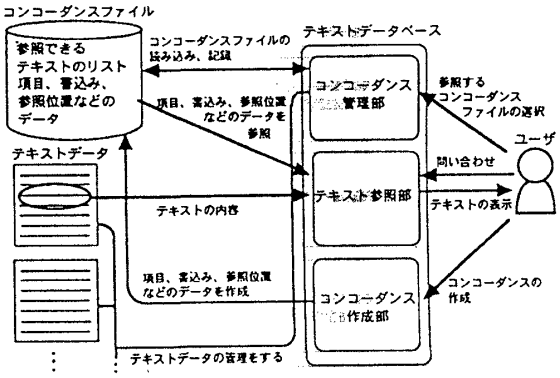

図１　システムの全体構成

データは、次の二つからなる。

(1) コンコーダンスファイル　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、書込みなどデータ

(2) テキストデータ　コンコーダンスの作成、参照対象のテキスト

3.2 システムの機能

機能を作業の流れに沿って説明する。

(1) 参照するテキストを管理する。

コンコーダンスは、一つの分野に関して複数のテキストから作成される。そのため参照するテキストをあらかじめコンコーダンスに登録する。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テキストを登録する。これをコンコーダンス管理部でおこなう。

Design and Implementation of a Text-Database Pointing to a Concordance
Motohiko SAKAGUCHI , Eiichi HAYAKAWA, Mitarou NAMIKI and Nobumasa TAKAHASHI
Faculty of Technology , Tokyo University of Agriculture and Technology

(2) コンコーダンスを作成する。

まずテキストビューアで項目とする語句を探す。次に項目の参照位置を文字列検索を用いて登録する。例えば、歴史文献では同一人物に様々な表記をし、一つの項目で参照できた方がよい。そのため項目を作成時に、項目名と参照位置の検索条件を別に入力する。参照位置登録は三通りの方法で行う。

a. 自動作成 検索条件で自動的に登録する

b. 半自動作成 作成者が検索結果の登録を判定する

c. 手動作成 参照位置を作成者が手動で登録する

登録後に関連項目を作成し、項目に関する知識を書き込む。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。

テキスト参照部での過程を図2に示す。

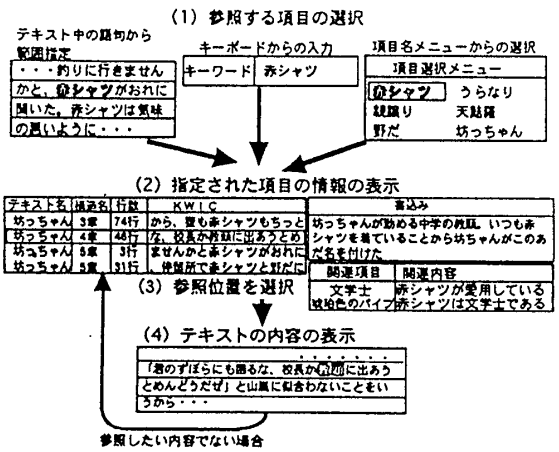

図2 テキスト参照の流れ

3.2 コンコーダンスの設計

例として夏目漱石『坊っちゃん』の「赤シャツ」の項目を図3に示し、各部分について説明する。

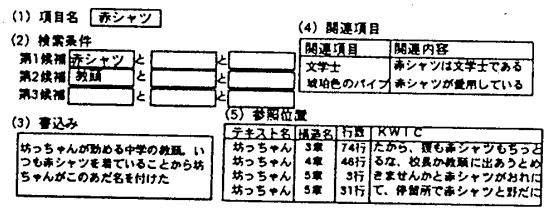

図3 コンコーダンスの項目の例

(1) 項目名 項目のラベルである。

(2) 検索条件 参照位置登録の文字列検索条件で、候補はOR検索、「と」はAND検索に対応する。

(3) 書込み 項目に対する書込みである。

(4) 関連項目 項目に関連する項目である。どんな関連かを関連内容で表す。

(5) 参照位置 参照する位置で、次の情報を持つ。

a. テキスト名 参照位置があるテキスト名

b. 構造 参照位置の章、節などの論理構造

c. 行数 参照位置が存在する行数

d. KWIC 参照位置のKWIC

4 システムの実現

システムの初版を、NEC PC−98 のMS−DOS上でC言語を用いて実現した。ソースは約8000行である。

5 『歴代宝案』の読下し文コンコーダンス

琉球外交文書『歴代宝案』の中の100通の読下し文にコンコーダンスを作成した。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

作成に約9時間かかった。今回自動化したのは語句の探索と、KWICの作成だけであるがかなり作成時間が短縮されたと考えられる。

次に本システムの使用例を述べる。『歴代宝案』でパレンバンは、「旧港」「宝安邦」「宝林邦」と3種類の表記がされる。項目に「旧港」と「宝安邦」を作成し、関連項目とした。「宝林邦」は、「宝安邦」の検索条件の第二候補とした。そのため使用者はどの表記からも参照できる。図4に実行画面を示す。

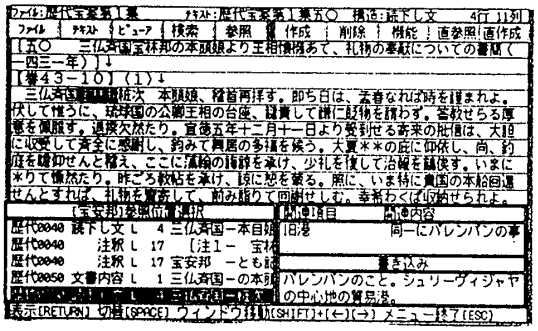

図4 実行画面

作業をおこなって得た知見を次に述べる。

(1) 研究対象への知識をまとめるのに役立つ

著者らは『歴代宝案』に対して知識がなかったが、調べて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コンコーダンスの作成を通して、『歴代宝案』の理解に役立った。

(2) テキストからの直接参照が有効である

『歴代宝案』のように、通常使用しない語句、漢字が多い文書では、キー入力は手間がかかり、テキストから語句を指定しての検索は有効であった。

6 おわりに

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。今後の研究課題としては、イメージとのリンク、項目の切り出しの自動化などが上げられる。

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