2003－CH－60　(5)
2003／10／24

Ｗｅｂを利用した歴史史料英日全文連携検索システムの開発

-　日米共同研究について　-

桶谷猪久夫[*1]　Delmer Brown[*2]　藤本雅彦[*1]　大久保祐子[*2]
[*1]大阪国際大学人間科学部、[*2]University of California, Berkeley

歴史史料を対象に、その文書構造や歴史的記述方法に着目し設計された英日全文連携検索システムを開発し、インターネット上に公開することにより、歴史学研究を援用し、さらに、国際的なコラボレーションを促進する。

直接対象とする文献は、日本の記紀である「古事記」、「日本書紀」や「続日本紀」、神祇関係の法令である「延喜式」、特定の地方誌的文書である「出雲国風土記」、日本初の解釈歴史書である「愚管抄」等であり、さらに、日本古典文献 25 巻のディジタル化を目標にしている。

本システムの開発と研究の目的は、英語を話す研究者や学生の日本史・国文学の研究に貢献することであり、また、日本の研究者との共同研究を促進することで研究の相乗的な効果を追求することである。

そのため Web 上で英語と日本語（または、両言語)を利用した文献内検索と文献間連携検索機能と閲覧機能を実現した。また、外字属性データベースを作成し、それを利用した歴史史料検索システムでの外字検索機能（入力法含む)、外字表示・転送機能を開発・実現した。

A Collaborative Research between Japan and the U.S. on Developing the Full Text Coordinated Retrieval System of Japanese Historical Documents using an Internet Website

Ikuo Oketani[*1]　Delmer Brown[*2]　Masahiko Fujimoto[*1]　Yuko Okubo[*2]
[*1]: Faculty of Human Sciences, Osaka International University
[*2]: University of California, Berkeley

This project develops the interactive retrieval system English and Japanese texts designed with a focus on language structures and historically descriptive methods. Furthermore, it promotes international collaboration by making these interactive retrieval system available to the public through the Internet, and by giving assistance to historical research.

The literature we are dealing with is as follows: 1) Japanese ancient chronologies such as *Kojiki*, *Nihon Shoki*, and *Shoku Nihongi,* 2) *Engishiki*, a collection of laws and regulations on shrines, 3) *Izumo Fudoki*, a local document of a certain area, 4) *Gukansho*, the first interpretive history text in Japan, 5) *Manyoshu*, an anthology, and so on. Our goal is to digitize twenty-five volumes of Japanese classic literature.

The goals of the development of this system and this research are two-fold: 1) to contribute to the research of Japanese history and literature conducted by English-speaking researchers and students, and 2) to pursue the synergistic effect of research by promoting collaboration between the English-speaking researchers and Japanese researchers.

With these goals, we constructed 1) a search program within texts using English and/or Japanese, 2) functions connecting and displaying different texts on the website. We also created 3) an attribute database of *gaiji* (non-standard *kanji* characters), and 4) a search function for *gaiji* (including input methods) for historical materials using the former database, and developed 5) functions to display and forward *gaiji* on the Internet.

１．はじめに

インターネット技術の発達、急速な普及と記憶容量の大容量化というこの革新的な変化は、コンピュータの利用を自然科学の分野から社会・人文科学の分野にまで急速に拡大し、それらを利用した電子情報の公開が一般的になってきている。ユーザーが必要とする情報を高速・的確・効率的に見つけ出すための情報検索システムの研究開発が必要不可欠になってきている。これは歴史学研究分野においても例外でなく、古典文献を電子化し、研究に活用しょうとする動きが盛んになりつつある。つまり、WWW(World Wide Web, 以下、Web という)上で提供される豊富な電子情報を有効に活用することで、創造的な活動の活性化が大いに期待される。

このような状況下で、我々は歴史史料を対象に、その文書構造や歴史的記述方法に着目し設計された英日全文連携検索システムを開発し、インターネット上に公開することにより、歴史学研究を援用し、さらに、国際的なコラボレーションを促進することを目的とする。

我々は、本システムの設計と開発をカリフォルニア大学バークレー校を中心に米国内外の研究・教育機関などによる研究プロジェクトで、学術研究と国際的コラボレーションを促進し、歴史史料のディジタル化と時間軸（年代）を設定できる地理情報システムとの連携を目指している ECAI (Electronic Cultural Atlas Initiative)と共同で開発した。その具体的な研究・開発プロジェクトは、ECAI 内の日本史研究を行っている JHTI (Japanese Historical Text Initiative)プロジェクトであり、日本古典文献 **25** 巻（後述)のディジタル化とデータベース化を目標にしている。

本稿では、歴史史料の日本文、英訳文とページイメージ画像ファイルに対して、連携して検索を可能にするため、それら各文書に対してデータ記述の定義（簡易型タグ付け)を作成した。その簡易型タグ付けをされた文書に対して、検索機能を設計し実現した。まず、本検索システムの目的と概要、各種検索機能、英訳支援システムとその問題点について述べる。また、古典文献を対象に検索システムを開発するとき必ず問題になる外字を含む文字列の入力、検索、表示、インターネット上での転送の解決策について述べる。

２．歴史史料英日全文連携検索システムの目的と対象文献の概要

本検索システムの目的は、特に日本神道を中心に古代日本の文化と日本人の精神生活の研究、その当時の事物や社会の様相を研究する資料を提供することにより、日本文化の世界への発信と国際的なコラボレーションを促進する研究である。また、外国人研究者の古典入門や研究支援だけでなく日本に関する教育にも役立つと思われる。さらに、歴史学研究に新たな視点を与え、新しい研究課題・方法を生み出す契機となり、コンピュータ応用技術やインターネット利用技術を新たな段階へ進展させる意義を持つと思われる。そのため歴史史料を対象にこれまで開発し公開してきた検索システム（基礎的実験も含む)[1,2,3〕 をさらに改良し、以下のことを実現する。

(1) 歴史史料の文書構造と歴史的記述方法に着目した検索手法を開発し、システムを設計・開発する。
(2) 関連する歴史史料の横断的検索機能の開発と歴史的変遷を考慮した履歴データベースを開発する。
(3) 歴史史料の定量的解析の試みと歴史的事例に基づく各種検索機能プログラムを開発する。
(4) 外字を対象とした漢字データベースの拡張とインターネット上での利用技術を開発する。
(5) 文献情報と古典史料を取り扱うとき重要な地理情報との連携化を実現する。

直接対象とした文献は、日本の記紀である「古事記」、「日本書紀」、「続日本紀」や「神皇正統記」、神祇関係の法令である「延喜式」、特定の地方誌的文書である「出雲国風土記」、歌集「万葉集」、中世の代表的歴史書であり全 7 巻から構成される「愚管抄」である。さらに、後述の日本古典文献 **25** 巻のディジタル化、**Web** 上で英語と日本語（または、両言語)を利用した文献内検索と文献間連携検索、閲覧と再利用を目標にしている。これら対象とする文献の一部は、既に研究者によりフルテキストとして、計算機可読の形式で入力済みで研究整備がされている。

３．歴史史料英日全文連携検索システムの設計・構築と各種機能の概要

本検索システムの開発と研究の目的は、英語を話す研究者や学生の日本史・国文学の研究に貢献することであり、また、日本の研究者との共同研究を促進することで研究の相乗的な効果を追求することで

ある。英語圏と日本の歴史研究者が、日本に関する歴史文献を共同研究することは重要である。その研究は、日本の古代史研究、日本古代国家の成立史や構造の研究、民俗（民族）学的研究であり、日本からの情報発信の先駆けになると思われる。

電子化情報の特徴として、検索、加工、複写、転送が容易であり、また、統計的処理やデータベース処理が可能であることなどがあげられる。しかし、歴史学研究で利用される古典史料のデータベース化や情報検索においては、歴史的に関連ある史料の効果的な横断的(統合的)検索機能の実現、外字や異体字の問題、原テキストの入力方法、出力方法など解決すべき種々の問題が存在し、いまだに有効な手法がないのが現状である。しかし、世界的規模での情報検索と情報発信が可能になったインターネット上の WWW(World Wide Web)を利用した研究は、歴史学研究分野においても、例外なく急速に普及している。

そこで我々は、本検索システムを近年の有力な研究基盤となっている Web 上で設計し構築した。つまり、Web 上で日本語と英語（または、両言語）を利用した文献内検索と文献間連携検索、閲覧、再利用を目標に実現した。英語圏と日本語圏の研究者が、歴史学研究に有効な史料検索システムを利用し、研究を進めるには、日本語文書と英訳文書が連携して、検索可能にならなければならない。そのため、４種類の文献、つまり日本語文書、英訳文書、ローマ字読み文書、原本に近い底本の画像ファイルに対して、文書構造が定義可能な簡易型のタグ付けを行った。そのタグ付けされた４種類の文書ファイルが連携して検索可能になる。簡易型タグについては、データ量が膨大になったとき、全文からの単純なパターンマッチング技法だけでは検索効率を考慮したとき問題があり、また検索条件を適切に指定できず効率的な検索には大きな制約がある。さらに、大量のデータから利用者の所望のデータを高速にかつ効率的に検索するには、全文検索システムが必要になる。この問題を考慮し、文書ファイルのデータベース管理システムへの格納とタグの拡張、例えば文書の論理構造を定義可能なマークアップ言語 XML (extensible Markup Language)への自動変換を前提にタグの設計を行った。簡易タグ付けの基本は、(1)既にディジタル情報として入力された英訳版を元にタグ付けする、(2)検索と表示の単位は、パラグラフ単位とする。そのため、漢文や英文は複数文になる場合があるので、それら複数文を 1 つのタグで囲む。タグ付けは、結果としてタグ個数が少なく、また簡易になり作業日程を短縮できた。

本検索システムは、現在 Web 上の CGI (Common Gateway Interface)機能を利用しインタープリタ言語 Perl (Practical Extraction and Report Language)で各種検索機能を実現している。

以下に、紙面の都合上最初に開発した江戸時代の儒学者で尾張藩士、河村秀根、益根父子が 60 年の月日を費やし刊行した「日本書紀」の注釈書である「書紀集解」[5]を例に各種機能を説明する。

3.1 歴史史料英日全文連携検索システムの各種検索機能

(1) キーワード検索機能

本検索システムが対象にした文献は、冊子体で和文と漢文で記述され非分割語で構成されている。そのため、検索システム構築の初期段階では、コンピュータによるキーワードの自動抽出は困難である。現在は、CGI 機能を利用しプログラムで、Web からの利用者の要求（文字列やその論理結合質問)を解釈し、格納されたデータに対して検索、つまり適切な文書の部分をパターンマッチングして取り出し、見やすく加工して表示している。また、指定されたキーワード(文字列）をログファイルとして蓄積し、再利用可能にした。キーワード検索機能について、以下に「書紀集解」の具体的な検索例で説明する。まず、図１で示すバークレーの JHTI (Japanese Historical Text Initiative)プロジェクトのホームページ[7]のプルダウンメニューから検索する文献を選択し、本検索システムを実行する。図２に、入力フォーム画面(Interactive Searching of Nihon Shoki)から「日本書紀」の検索対象巻番号を選択する。次に、MODE : ボックスで Retrieval を選択し、Find word or phrase: ボックスに、記紀神話における男神の伊弉諾尊（いざなきのみこと）と女神の伊弉冉尊（いざなみのみこと）が大八洲國を誕生（国生み）させ、次に海、川、山を生み、日神（ひのかみ）を生む箇所から、「大日%u5b41;」を入力し、Word(s) retrieval : Which version? ボックスで **Japanese** を指定し検索した例を示す。(注）%u5b41;は外字（図３と４で外字を参照可能）

図３に、検索結果が表示される。画面の上部から、検索対象文献名(Nihon-shoki)、検索キーワード(大

日%u5b41;)、文献の巻数名("巻第１（神代上）- １. THE AGE OF THE GODS I" "巻第２（神代下）- ２. THE AGE OF THE GODS II")とマッチしたパラグラフ数(Found : 3 matches)と該当するパラグラフ一覧がページ数とパラグラフ番号と共に表示される。また、指定したキーワードは、見やすくするため赤色太字で表示される。詳細表示（英日対応文書)を希望するとき、該当パラグラフのMore Details...をクリックすると日本語と英語の対応パラグラフが表示される。パラグラフ数１の場合は、図４に示すように、表示画面の右側に、「書紀集解」の画像ファイルから切り出された該当文字列が画像で表示される。画面上部の[This Page's Image]をクリックすると、Original Image of Document の該当ページ画像が表示される。また、キーワード前後の文書が参照できるように、表示パラグラフ数を３または５に変更可能である。図５は、表示パラグラフ数が５の表示例である。さらに、英訳文書においては[Show Notes] をクリックすることにより、本文に続きノート（注釈行）が該当箇所に展開表示可能である。

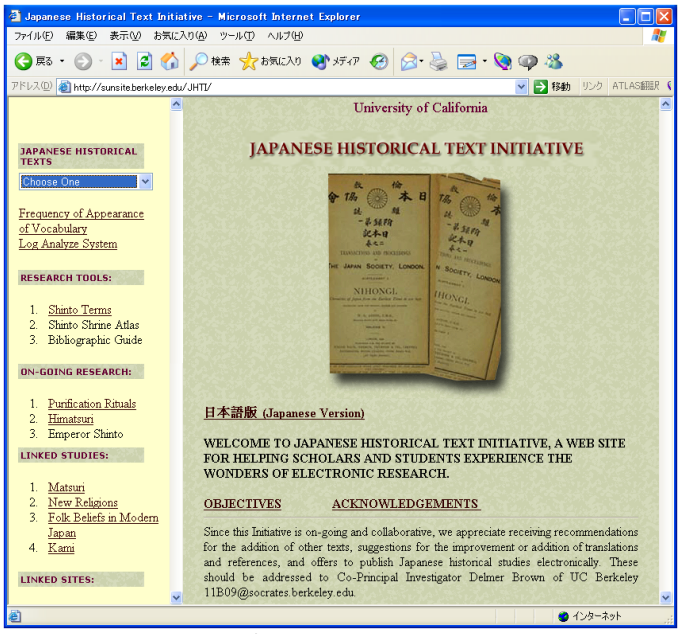

図１．JHTI プロジェクトのホームページ

(2) 項目検索機能

この機能は、神の名前、神社名、神社の場所名、儀式などから文献を効率的に検索することを想定している。この機能は、現在、キーワード検索と同様な機能しか有していないが、今後の拡張で XML タグなどを付加したときに有効に作用すると思われる。そのため、各項目のテーブルを作成し、タグ付けをプログラムで自動的に行った。

(3) 閲覧(ブラウジング）機能

この機能は、言語（英語、日本語、両言語対応）を選択し、文献を巻番号（複数巻指定可）の先頭から、またページ指定やパラグラフ番号指定で、連続して閲覧することが可能である。さらに、閲覧するパラグラフ数を指定可能である（5,10,20,30 デフォルト値：10)。閲覧(ブラウジング）機能は、日本史・国文学を学習

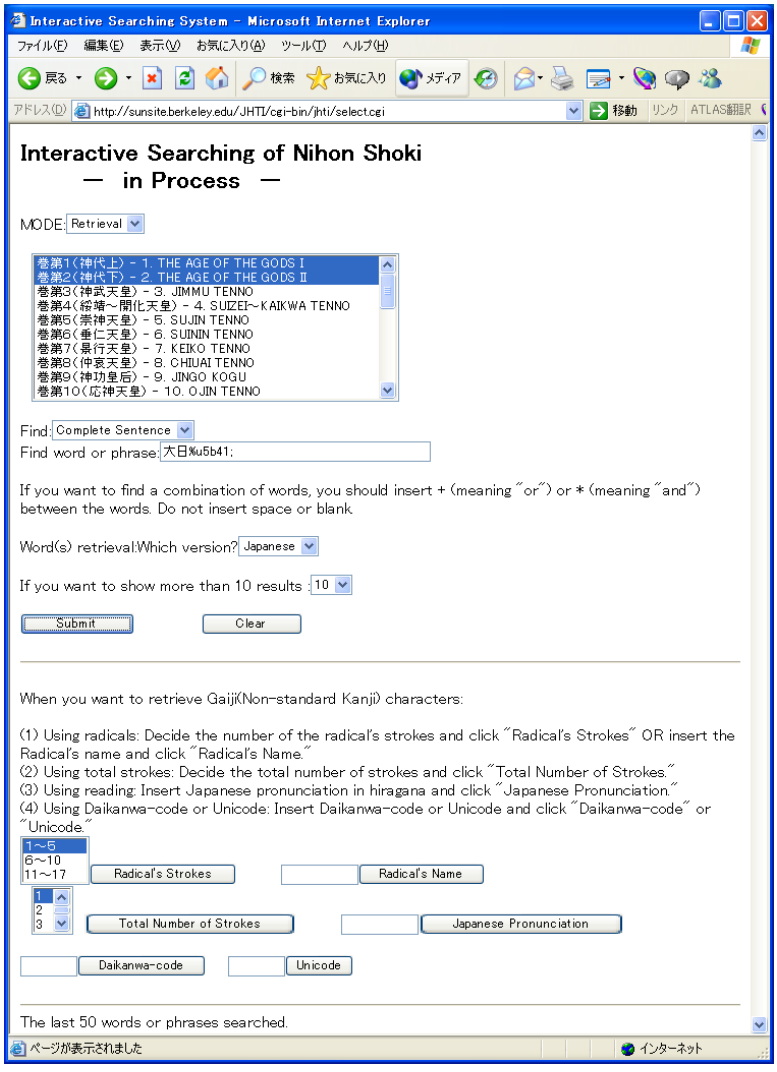

図２．入力画面でキーワード(大日%u5b41;)を指定した例

する初心者にとって、また、それらの教育用に有効であると思われる。

(4) 史料の定量的解析の実験

電子化情報の特徴として、検索、加工、複写、転送が容易であり、また、統計的処理や計量言語学などの手法を利用した研究の展開の可能性を持っている。今回、JHTI プロジェクトのホームページのメニュー画面で Frequency of Appearance of Vocabulary をクリックすることで、巻単位、文献単位や複数文献単位に単語（Word）の使用頻度の調査を可能にした。

本歴史史料英日全文連携検索システムは、研究・教育に試験的に公開され使用されている。開発における問題点として、①単語の表記上の違いがあった。たとえば「古事記」の英訳本では、男神伊弉諾尊は、”IZANAGI”となり、「日本書紀」の英訳本では、”Izanagi”となっている。これらの問題は、検索用プログラムで対処(解決)した。②既にディジタル化された英訳本の注釈(NOTE)が冊子体のイメージに忠実に入力されていた。この問題に対しては、検索プログラムのバッファリングで対処しても検索効率の低下をもたらすため、注釈(NOTE)を別ファイルとして保存し、本文の位置情報を示すタグで対処した。③各文献で研究者に有効な検索方法の採用が必要である。例えば、「続日本紀」は、編年体の史書の特徴として、巻番号の後に年・季・月・日が記述されている。そのため検索機能としてキーワードのみでなく年月日や指定期間内の検索も可能にした。

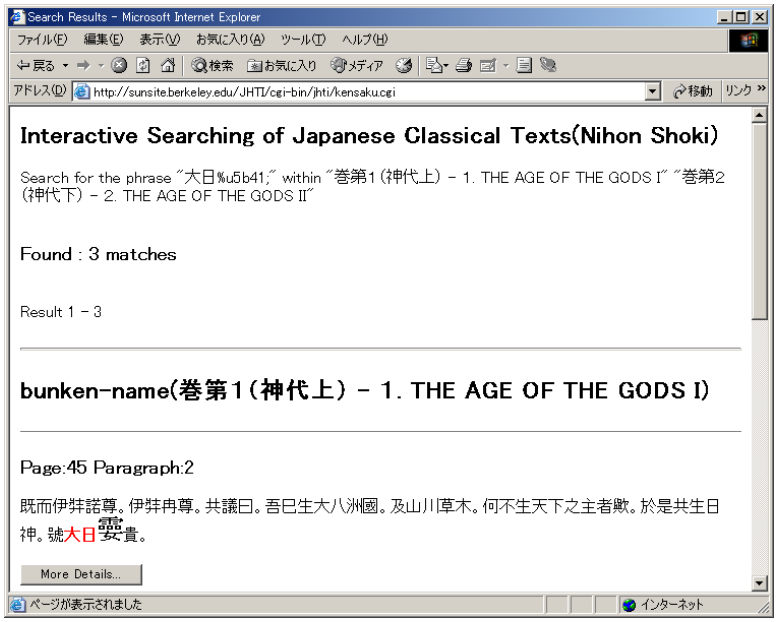

図３．歴史史料英日全文連携検索システムの検索結果画面

図４．英日対応パラグラフの表示例

3.2 日米共同研究の推進

本歴史史料英日全文連携検索システムは、Web 上で日本語と英語（または、両言語）を利用した文献内

検索と文献間連携検索、閲覧、再利用を可能にする。そのため、後述の 25 文献の英日ディジタル情報が必須になる。このため、現在、国文学研究資館の科学研究費補助金（基盤研究Ｓ)「国際コラボレーションによる日本文学研究資料情報の組織化と発信」と「JHTI(カリフォルニア大学バークレー校)プロジェクト」との共同研究を目指して協定書締結を準備している。これにより、JHTI プロジェクトが保有していない日本古典文学作品データベース（日本古典文学大系）のディジタル情報を使用し研究を遂行可能になる。当然、この部分の使用は国文学研究資料館プロジェクトとの協定書に許可された範囲内であり、また国文学研究資料館に登録された日本文学・史学研究者に限定される。現在、この協定書を前提に、「神皇正統記」や「大鏡」等の開発を進めている。認証システムの概念図を図６に示す。

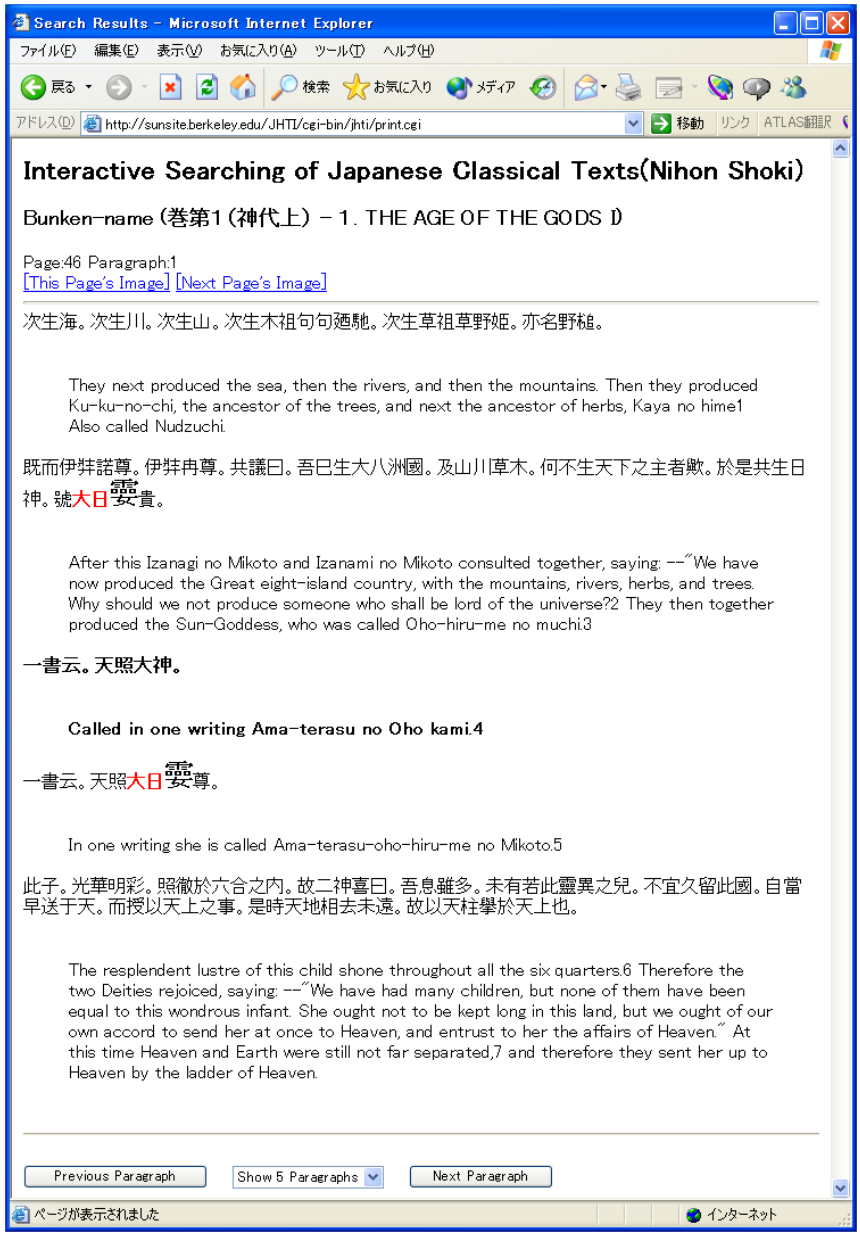

図５．パラグラフ数５の詳細表示画面

3.3 英訳（翻訳）支援システムの実現

「続日本紀」は、文武天皇元年(697)から桓武天皇の延暦十年(791)まで、九代 95 年間の歴史を記述した漢文の史書である。「日本書紀」に次ぐ第２の正史として編纂され延暦 16 年に 40 巻の書として完成した。古代の歴史・国文学を研究する上で重要な文献である。

しかし、本歴史史料英日全文連携検索システムが対象にした「続日本紀」の英訳本(J.B. Snellen)[6] は、巻一から巻六までしか完成していないため、それ以降の英訳の完成が求められている。それらの要望に答えて、Web ベースの英訳支援システムを実現した。「続日本紀」の編年体の史書の特徴を生かし、年・季・月・日を自動的に挿入している。また、既に検索可能になっている文献（古事記、日本書紀など）の参照も可能にしている。その翻訳支援システムの入力画面を図７に示す。現在、この支援システムを利用し第六巻以降の翻訳が遂行されている。

４．歴史史料英日全文連携検索システムの外字処理機能

古典文献を対象に、文書検索システムを構築するとき、外字や異体字の問題を解決することが不可欠である。外字に対する入出力や検索機能の効果的な実現法が存在していないのが現状である。しかし、研究

者はできるだけ原典に近い形式で研究を遂行したいという要望や、外字や異体字そのものを、また、それら文字の文献の文脈中での使われ方そのものを研究対象としている。

本歴史史料英日全文連携検索システムは、インターネット環境下でのテキスト情報の検索サービスを提供するため、(1)外字の入力方法、(2)外字を含んだ文字列検索、(3)外字を含んだ文字列表示、(4)外字の転送方法、を解決した。表１に、「古事記」、「日本書紀」、「延喜式」、「続日本紀」に出現した外字数と外字種類を示す。以下に、外字処理における入力方法、検索手法、出力（表示）方法、転送方法について簡単に述べる。

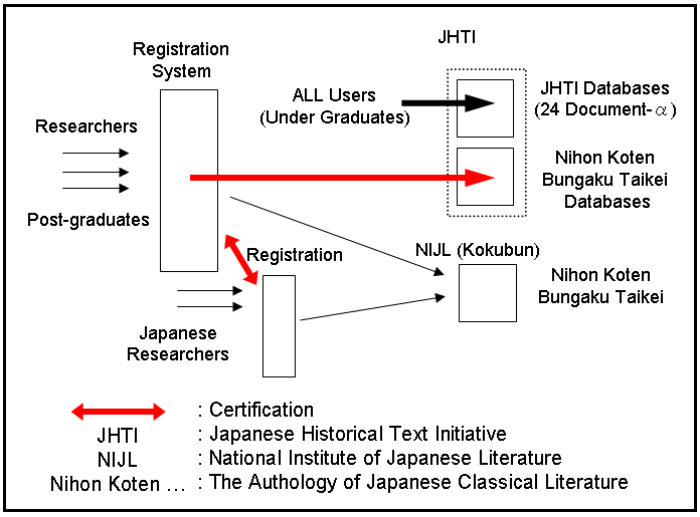

図６．JHTI 認証システムの概念図

(1) 外字の入力法

外字の入力については、%u と;（一種のタグの役割）で囲んで入力、Unicode に存在するフォントはそのコードを入力し、存在しない漢字に対しては、%uxxxx；の xxxx を f0001 から順にコードを割り振って入力した。

(2) 外字検索機能、

外字を含む文字列や外字の検索には、(a)部首の画数選択、(b)直接部首名の入力、(c)総画数の入力、(d)音読みの入力、(e)大漢和コード入力、(f)Unicode 入力で外字を選択可能である。

(3) 外字表示機能と外字転送機能

Web 上での外字の表示機能と転送機能は、画像ファイル(GIF 形式ファイル)の張り付けと転送で解決した。

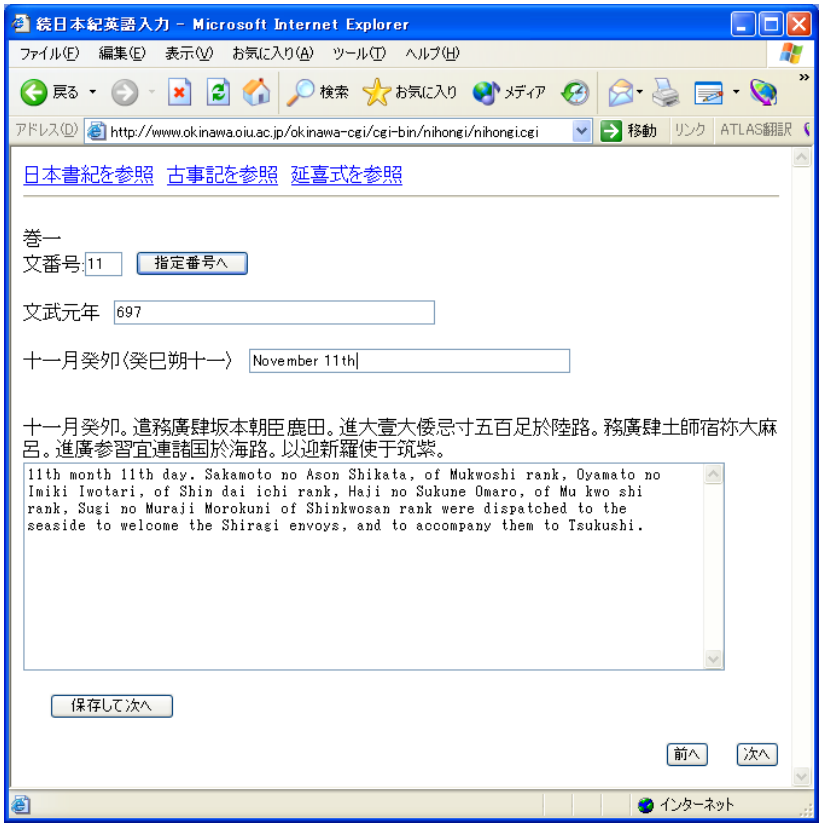

図 7. 英訳（翻訳）支援システムの入力画面

５．おわりに

本稿では、「日本書紀」を題材に、インターネット上の Web による複数文献ファイル、例えば日本語ファイル、英訳ファイルと画像ファイルの英日全文連携検索システムの実現、各種検索機能と外字処理について述べた。本検索システムはインターネット環境下で、複数文献のテキスト連携表示機能、ページ画像連携表示機能、外字の混在した文字列の検索機能と外字表示機能／転送機能に対して、有効に作用している。本検索システムが日本の史料の新たな解釈・解析など歴史学研究を促進し、研究支援へのコンピュータの有効性を示し、新しい視点を与え、新しい研究課題と研究方法を生み出す契機になっていくことを期待したい。今後、表２に示す文献の格納も計画している。また、データベース管理システム(OpenText を想定)への格納、タグとして XML への機能拡張を早急に実現したい。

表 1. 各文献に出現した外字数と外字種類

	評定所文書	日本書紀	古事記	延喜式	続日本紀
外字総数	7,512	969	545	1,340	153
ユニコード内	1,995	838	210	1,222	142
作成外字	5,517	131	235	118	11
内大漢和内	119	83	227	32	7
内大漢和外	5,398	48	8	86	4
外字種類	317	305	83	109	54
ユニコード内	201	230	54	94	55
作成外字	116	75	29	15	7
内大漢和内	28	44	23	10	5
内大漢和外	88	31	6	5	2

最後に、本検索システム構築の機会を与えてくれたカリフォルニア大学バークレー校の ECAI コーディネータ Lewis Lancaster 教授、歴史史料に対するご教示やご討論を頂いた東アジア図書館(East Asian Library)図書館長 Peter Zhou 氏、ライブリアン石松久幸氏、文献の英訳文書と日本語文書の校正・編集をやっていただいた新谷廣一氏ほか関係各位に謝意を表する。なお、本研究は科学研究費基盤研究(B)(2)「Web を利用した歴史史料英日全文連携検索システムの設計と開発に関する研究」(平成 15～17 年度、研究代表者　桶谷猪久夫）の下で行った。

表 2．ディジタル化対象文献（*検索可能、**タグ付け中）

Text 1: Kojiki (古事記) *
Text 2: Nihon Shoki (日本書紀) *
Text 3: Shoku Nihongi (続日本紀) *
Text 4: Izumo Fudoki (出雲風土記) **
Text 5: Kogoshui (古語拾遺)
Text 6: Engi Shiki (延喜式) *
Text 7: Eiga Monogatari (栄華物語)
Text 8: Okagami (大鏡) **
Text 9: Azuma Kagami (吾妻鏡)
Text 10: Gukansho (愚管抄) **
Text 11: Jinno Shotoki (神皇正統記) *
Text 12: Taiheiki (太平記)
Text 13: Daijingu Jin'iki (大神宮神威記)
Text 14: Dokushi Yoron (読史余論) **
Text 15: Meiji igo Shukyo kankei Horei * (明治以降神社関係法令史料)
Text 16: Kokutai no Hongi (国体の本義) **
Text 17: Tenri-kyo (天理教)
Text 18: Kurozumi-kyo (黒住教)
Text 19: Konko-kyo (金光教)
Text 20: Omoto-kyo (大本教)
Text 21: Itto-en (一燈園) **
Text 22: Tensho Kotai Jingu-kyo (天照皇太神宮教)
Text 23: Rissho Kosei-kai (立正佼成会)
Text 24:Tsubaki Ookami Yashiro (椿大神社)
Text 25:Manyousyu (万葉集)

【参考文献】

[1] 桶谷猪久夫、新谷廣一『SGML を利用した琉球王国評定所文書と琉球家譜の全文連携検索システムの設計と実現』、大阪国際女子大学紀要 27 号-2, pp. 1-18, 2002.3.31
[2] Ikuo Oketani, Chizuko Saito, Delmer Brown “A Design and Construction of the Full Text Retrieval System using Simple-tagged Nihon-shoki Texts (the Imperial Chronicle of Japan)” ECAI (Electronic Cultural Atlas Initiative) Conference, Sydney, pp. 12-12, June 13, 2001
[3] Ikuo Oketani, Chizuko Saito, Delmer Brown “The Construction and the Future Development of the Full Text Coordinated Retrieval System of Historical Documents using the Internet” PNC (Pacific Neighborhood Consortium) Interim Conference, Guadalajara (Mexico), pp. 11 - 11, December 2, 2001
[4] Norinaga Motoori “Kokun Kojiki teisei 4-koku” Nagata Bunshodo, 1874, stored in UCB East Asian Library
[5] W. G. Aston ”NIHONGI: Chronicles of Japan from the Earliest times to A.D. 697” Printed by the Japan Society, 1896
[6] J.B. Snellen “Shoku Nihongi : Chronicles of Japan, continued, from 697-791 A.D.” In The Transactions of the Asiatic Society of Japan. (Vol.1-3) Second Series. Vol.11. 1934. Asiatic Society of Japan; 151-239, (Vol.4-6) Second Series. Vol.14. 1937. Asiatic Society of Japan; 209-278
[7] Japanese Historical Text Initiative : http://sunsite.berkeley.edu/JHTI/