様式 C-19

科学研究費補助金研究成果報告書

平成 22 年　6 月　8 日現在

研究種目：基盤研究(A)
研究期間：2007～2009
課題番号：19202016
研究課題名（和文）　日韓言語横断歴史資料検索システムの研究

研究課題名（英文）　Study on cross-language retrieval system for Japanese and Korean Historical Terms

研究代表者
鶴田　啓（TSURUTA KEI）
東京大学・史料編纂所・教授
研究者番号：10172066

研究成果の概要（和文）：東京大学史料編纂所にサーバを設置し、1)史料編纂所の宗家史料データ全件、2)長崎県立対馬歴史民俗資料館の宗家文庫史料の一部、3)国立国会図書館所蔵の宗家文書データ全件についてマッピングを検討した上で搭載し、史料編纂所内および対馬歴史民俗資料館から検索可能とした。また、言語横断検索については、複数の方法により歴史熟語と固有名詞等について日本語・ハングルの対訳辞書を作成し、検索語がハングルの場合は辞書による検索をはたらかせ、1)歴史用語辞書中に語が存在する場合はその対訳日本語で検索し、2)存在しない場合は別字書により漢字文字列へ変換して検索を行う機能を作成した。

研究成果の概要（英文）：We researched about uniting various format catalogs and installed the following data about So family's historical documents on a server settled in Historiographical Institute Univ. of Tokyo(HI). 1)all data of owned by HI, 2)sample data provided from Nagasaki prefectural museum of Tsushima history and folklore, and 3)the National Diet Library(NDL)'s data opened to the public. Then we made possible to retrieve from the museum of Tsushima history and folklore. About cross-language retrieve, we made the bilingual dictionary about history idioms and proper nouns between Japanese and Korean by some methods and designed the following retrieve function. When retrieval word was input in Hangul characters, 1)if the word exists in the historical dictionary, retrieve by the Japanese translation, 2)if it did not exist in the dictionary, change the word into Chinese characters(KANJI) by another dictionary and retrieve by KANJI.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
19 年度	9,800,000	2,940,000	12,740,000
20 年度	8,500,000	2,550,000	11,050,000
21 年度	6,400,000	1,920,000	8,320,000
年度			
年度			
総　計	24,700,000	7,410,000	32,110,000

研究分野：史学
科研費の分科・細目：史学一般
キーワード：文化交流史

様式 C-19

科学研究費補助金研究成果報告書

１．研究開始当初の背景

(1)インターネットの普及に伴い、歴史資料（史料）を対象とする国際的な検索は飛躍的に増大し、史料編纂所が WEB 上に公開している歴史資料データベースも、国外からの検索を多数受け付けている。一方で、データベースの分野では内部処理が国際的に互換性のある文字コードセット（UTF-8）で行われるようになり、WEB ブラウザの多くが多言語表示に対応するようになって、通常の（現代文で表記された）WEB ページではボーダーレス化が進んだ。

(2)しかしながら、歴史の分野においてこれまで言語横断検索システムの研究および実運用システムはほとんど存在しなかった。その理由としては、歴史的な用語や概念は複雑な意味合いを持っており、従来からシソーラス（類義語辞書）の研究・作成はあったものの、単純に外国語に置き換えるだけでは情報互換にならないという問題があった。

(3)2006 年度に史料編纂所の付属施設として前近代日本史情報国際センターが 4 年時限で設置されたことにともない、ここを検索システムの開発研究を担う情報学と歴史学との接点と位置づけ、歴史情報の国際互換を研究の柱の一つとした。

２．研究の目的

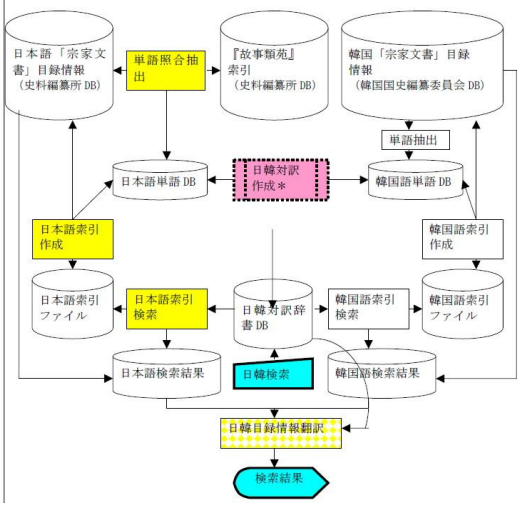
(1)漢字語レベルでの言語横断検索を実用的に実現する前処理と手法の研究。ある漢字（あるいは熟語）が、歴史上同じ意味で使われる場合には、文字コードの違いを処理するだけで対応が可能であるが、たとえば日本の江戸時代特有の用字用語、あるいは朝鮮王朝時代特有の用字用語については、解釈や置き換えが必要になる。史料編纂所では以前から日英、日仏などの言語を対象にシソーラスを作成してきており、その成果は「日本史グロッサリー・データベース」として利用に供しているが、その作成には高度な専門知識を持った研究者と多くの時間が必要であった。本研究では、対象を漢字語に限定し、自然言語処理の手法を用いてテキストを解析することで、より少ない時間と人手で言語横断検索を可能にする手法の検討を行う。

(2)歴史資料（史料）の横断検索に必要なマッピングの研究。歴史資料の目録化・データベース化は各機関がほとんど独自に記述を行っている。たとえば、かつてひとまとまりであった文書が現在複数の機関に分かれて所蔵される場合でも、目録書誌の採り方やそれを元に電子化されたデータは、異なるフォーマットを持っているのが普通である。このような独自性は、それぞれの歴史資料が持つ特性を各所蔵機関がどのように判断したかの表れであり、一概に否定されるべきものではない。とはいえ、一定の規則のもとで作成される刊行物の書誌目録データベース（例：図書館のオンラインカタログ）と比較して、横断検索を行うことを難しくしている。そこで、歴史資料における目録書誌データのマッピングについて研究を進める。

(3)マイクロフィルム画像スキャンによる歴史資料としての用途（解読）に適したデジタル画像作成。

(4)上記(1)の研究要件に基づき当初構想したシステム関係の概念図を[図１]に示す（原図作成：研究分担者石川徹也）。ただし研究の過程で、九州国立博物館・国立国会図書館・慶應義塾大学付属図書館などが所蔵する宗家文書のデータが各機関の公開データベースに載るようになって（慶應義塾大学付属図書館分は利用登録制）、上記(2)の要素がより重要性を増した。このため実施にあたり変更を加えた箇所がある。

図１日韓言語横断歴史史料検索システム機能概要



３．研究の方法

(1)言語横断検索の研究について

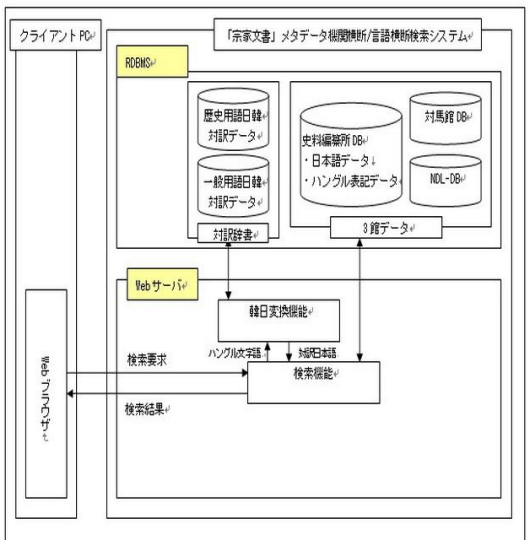
①日本の歴史資料に関する用語辞書の作成として、史料編纂所が所蔵する江戸時代対馬藩史料の目録データを自然言語処理の手法により解析し、同史料特有の用語の切り出しを行った。これを見出し語とし、日本語・韓国語の歴史用語を理解する研究者の手により韓国語（ハングル）の対訳を作成し、歴史専門用語の対訳辞書を作成した（辞書１）。

②朝鮮史で使われる用語（おもに漢字語）を見出し語とし、そのハングル表記（ただし第二次大戦前の表記法）と日本語による解説（同前）を付けた辞書である朝鮮総督府中枢院編『朝鮮語辞典』を入力し、研究者によるチェックを経て、歴史的漢字語とハングル（現代表記）の対訳辞書を作成した（辞書２）。

③検索語にハングルが入力された場合にデ

ータベース側でどのような処理を行うかについて、ルーチンとレスポンスを中心に比較検討した。
(2)機関間横断のためのマッピングについては、東京大学史料編纂所・対馬歴史民俗資料館・国立国会図書館それぞれが作成した目録データを入手し比較対照した。科研データベースへ搭載するにあたっては、元データの項目変更は行わず、記述内容が同一もしくは近似と見なしうる項目どうしを同一項目と見なして検索と結果表示をすることにした。
対象史料群の概要は次の通りである。
①宗家史料　東京大学史料編纂所が所蔵する宗家史料。大半が冊子形態、一部一紙形態。
②宗家文庫史料（冊子類）　対馬歴史民俗資料館が所蔵する宗家文庫史料のうち冊子形態の史料。
③宗家文庫史料（一紙類）　対馬歴史民俗資料館が所蔵する宗家文庫史料のうち一紙形態の史料。
④宗家文庫史料（絵図類）　対馬歴史民俗資料館が所蔵する宗家文庫史料のうち絵図形態の史料。
⑤宗家文書　国立国会図書館が所蔵する宗家文書で冊子形態。
(3)史料編纂所が所蔵する宗家史料の 35㎜ マイクロフィルム（171 リール）を対象にスキャンを行った。

４．研究成果
(1)言語横断検索機能については、検索語にハングルが入力された場合、辞書１＞辞書２＞辞書３（一般的な翻訳での訳語）の順で優先付けを行い、辞書１に該当語があればその訳語で検索、辞書１に該当語が無い場合辞書２を読み、該当語があればその訳語で検索、（以下同）という方法とし、検索対象項目に該当する語がある場合問題なくヒットすること、レスポンス的に問題がないことを確認した。作成したシステムの概要を次図に示す（原図作成：(株)ジー・サーチ）。



(2)機関間横断検索について、比較検討の対象とした各機関所蔵分データの概要は次の通りである。
①史料編纂所宗家文書目録、2944 件、史料編纂所フォーマット（冊単位）
②対馬歴史民俗資料館冊子類目録サンプルデータ、249 件、同館冊子類フォーマット（冊単位）
③対馬歴史民俗資料館一紙類目録サンプルデータ、205 件、同館一紙類フォーマット（一点単位）
⑤対馬歴史民俗資料館絵図類目録サンプルデータ、203 件、同館絵図類フォーマット（一点単位）
⑤国立国会図書館宗家文書目録データ、58 件、同館フォーマット（書目単位）
特に大きな違いは、史料編纂所と対馬歴史民俗資料館が物理的な１点を基本単位としているのに対し、国立国会図書館では書目を基本単位としていることである。検討の結果、マッピングは次表の通りとした。色つきの項目が横断検索の対象となる項目である。なお基本単位の違うデータを結果表示の段階で違和感なく見せることは今後の課題である。

本システム項目	史料編纂所	対馬(冊子)	対馬(一紙)	対馬(絵図)	国会図
書目ID	書目:書目ID	連番	通番	番号	請求記号
架-番-枝番	書目:架	所蔵管理番号1～4	調査番号	所蔵者管理番号	-
	書目:番				-
	書目:枝番				-
書名	書目:書名	史料名・表題	文書名	名称	タイトル
書名ヨミ	書目:書名ヨミ	-	-	-	タイトルよみ
作成	書目:著者名	作成・発給	作成	作成	-
出版事項	書目:出版事項1	-	-	-	-
形態	書目:形態1	形状	形態	形状	-
大きさ	書目:大きさ	法量	法量	法量	形態
注記1	書目:注記1	内容	備考	-	内容細目
注記2	冊:注記2	備考	備考	備考	-

備考	冊:備考2	-	-	-	-
年紀	細目:年紀	-	-	-	-
紙数 丁数	-	丁数	紙数↓丁数	-	-
ページ番号	-	ページ番号	-	-	-
作成年月日	書目:備考1	年月日	作成年月日	年月日	-
宛所	書目:著者名	充所	宛先	-	-
員数	書目:形態1	員数	員数	員数	-
注記3	-	-	-	-	注記
件名	-	-	-	-	件名
リンク	-	-	-	-	NDL-OPAC

(3)対象としたマイクロフィルムから約 10 万コマの JPEG 画像を生成し、史料編纂所の所蔵史料目録で画像が存在する宗家史料がヒットした場合にイメージ表示を可能とした。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）
①近藤成一、東京大学史料編纂所における横断検索システムの構築Ⅱ―非横断型システムによる研究情報資源連携の試み―、人間文化研究情報資源共有化研究会報告集、査読無、No. 1、2010、pp. 75-78
②渡邉正男、「宗家判物写」所載文書編年目録稿、東京大学史料編纂所研究紀要、査読無、No. 20、2010、pp. 1-24

〔学会発表〕（計４件）
①近藤成一、東京大学史料編纂所における横断検索システムの構築Ⅱ―非横断型システムによる研究情報資源連携の試み―、人間文化研究情報資源共有化研究会、2009 年 5 月 29 日、国文学研究資料館
②石川徹也、「歴史知識学の創成」研究、公開研究会「歴史知識学の創成」、2008 年 11 月 22 日、東京大学山上会館
③近藤成一、二一万通の古文書を集める―日本古文書ユニオンカタログプロジェクト―、公開研究会「歴史知識学の創成」、2008 年 11 月 22 日、東京大学山上会館
④遠藤基郎、鎌倉遺文を対象とする Virtual Laboratory 構築プロジェクト、公開研究会「歴史知識学の創成」、2008 年 11 月 22 日、東京大学山上会館

〔図書〕（計１件）
横山伊徳・石川徹也編、勉誠出版、歴史知識学ことはじめ、2009、202 ページ（石川徹也：pp. 1-15、遠藤基郎：pp. 81-99、近藤成一：pp. 101-117）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
鶴田　啓（TSURUTA KEI）
東京大学･史料編纂所・教授
研究者番号：10172066

(2)研究分担者
石川徹也（ISHIKAWA TETSUYA）
東京大学・史料編纂所・教授
研究者番号：20041808
近藤成一（KONDO SHIGEKAZU）
東京大学・史料編纂所・教授
研究者番号：90153717
榎原雅治（EBARA MASAHARU）
東京大学・史料編纂所・教授
研究者番号：40160379
遠藤基郎（ENDO MOTO-O）
東京大学・史料編纂所・准教授
研究者番号：40251475
箱石　大（HAKOISHI HIROSHI）
東京大学・史料編纂所・准教授
研究者番号：60251477
渡邉正男（WATANABE MASAO）
東京大学・史料編纂所・准教授
研究者番号：80230994
須田牧子（SUDA MAKIKO）
東京大学・史料編纂所・助教
研究者番号：60431798

(3)連携研究者　　(0 名)