様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年６月７日現在

機関番号：３２６３６
研究種目：若手研究（Ｂ）
研究期間：2010～2012
課題番号：２２７００２６２
研究課題名（和文）　実用的な明朝体漢字字形データベースの運用と応用

研究課題名（英文）　Application and operation of practical database of Mincho Chinese Character-shape

研究代表者
上地　宏一（KAMICHI KOICHI）
大東文化大学・外国語学部・講師
研究者番号：２０４６８７２１

研究成果の概要（和文）：漢字字形データベースの運用、対外的な公開およびデータの追加収録を行った。国際文字コード規格等で規定される全 9 万漢字・異体字を含む 31 万種の字形を収録するまでに至った。登録字形データを動的に呼び出し表示するための機能を構築し、様々な漢字・異体字を含む Web ドキュメントを容易に作成できることとなった。登録されている大量の異体字を利用時に弁別するための基礎データとなる「漢字異体化データベース」をまとめた。

研究成果の概要（英文）：I operated, published and added data of a Kanji-shaped database. The number of records of the database led to up to 310,000, including 90,000 kanjis and its variants defined by the international character code standard. I built a function to call dynamically display of registration shaped data, it can easily create a web documents that contain variety of Kanji variants. I published "Kanji variation database" which can be a basic data for discrimination when using a large number of Kanji variants in the database.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	500,000	150,000	650,000
２０１１年度	600,000	180,000	780,000
２０１２年度	700,000	210,000	910,000
年度			
年度			
総　計	1,800,000	540,000	2,340,000

研究分野：総合領域
科研費の分科・細目：情報学、図書館情報学・人文社会情報学
キーワード：文字データベース、異体字、Unicode、漢字字形

１．研究開始当初の背景

コンピュータ上での文字（漢字）処理は文字コードによって規定されており、日本国内では 1978 年に制定された JIS X 0208 規格がその後改訂を重ねながら 10 数年にわたって利用されてきた。だがコンピュータの普及・利用機会の増加により、当規格で規定される文字集合の範疇では足りない漢字の処理という問題が生じた。これに対しユーザー定義文字（外字）の利用によって解決することもあった一方、外字は情報交換を阻害する要因ともなり、近年のインターネット普及により実質的に外字の利用が難しくなった。

この問題に対していくつかのアプローチが提案・実用化された。ネットワーク上の閉じたコンピュータ同士で外字データを共有するアプローチや、工業規格とは別の私的な漢字集合を規定しフォントを販売・配布するアプローチが挙げられるが、工業規格（文字コード）に準拠して情報交換が保証され、か

つ、だれでも利用できる汎用的な解決法は出現しなかった。その後国際標準文字コード規格 ISO/IEC 10646（および当規格に相当する Unicode）が制定され、その後の拡張により７万字におよぶ漢字集合が収録されたため、「コンピュータで扱える漢字が足りない」という問題の多くは解消された。

ところが漢字は 7 万種類でも足りることはなく、人名などの細かい差異で区別が必要な異体字処理や、古典に見られる漢字や国字、学者によって字形解釈の異なる古代文字の扱いなど、文字コードの拡張では解決のできないコンピュータ上の漢字処理が問題となっている。このことは電子戸籍処理には法務省が独自に制定する文字コードが利用されていることや、電子出版・DTP の業界では人名や旧字体の印刷に適したメーカー独自の文字コードが普及している現状からも明白である。また大規模漢字コードの制定に対して実際に人が文字として利用するためにはフォントが必要であるが、現状ではすべての漢字を収録するフォントは日本には存在しないため、中国・台湾で制作された日本人にとってやや不自然なデザインのフォントを利用せざるを得ない。

あらゆる漢字を扱うための多漢字環境・漢字データベースの構築の関連研究として、国際的にも評価の高い「HNG 漢字字体データベース」の存在が知られているが、収集対象は写本や印刻本等に限定され、またデータをフォントとして利用することはできず、あくまで字形レコードを蓄積することが主眼となっている。また台湾の中央研究院においても古典籍の漢字を幅広く収録したデータベース・フォント集合である「漢字庫」が構築されているが、文字収集には当該機関の主観による採録基準が設けられているため、すべての文字を収録できるということではない。このほかに XML での利用を想定した、あらゆる文字・記号を登録する国際グリフ登記簿の運用を規定する国際標準 ISO/IEC 10036 も存在するが、利用についての具体的な規定はなく実用段階には至っていない。一方で多漢字環境の実現を阻害する要因として漢字フォントはその製作コストが高いこともあり財産物として自由利用に対する制限が課せられやすいことが挙げられる。その結果、現状では各ユーザー・機関が独自に外字集合としてのデータを保持するにとどまっており、散在している漢字情報を網羅し、ユーザーが自由に活用できる実用的な多漢字環境の実現が求められている。

２．研究の目的

研究代表者は明朝体漢字字形について、その字形を維持したまま極度に単純化するデータ形式の研究を過去に行ってきた。具体的には漢字の筆画を始点・終点といった位置座標情報と、筆画の種別・端部形状の集合で表現するモデルを考案した。さらに他の漢字部品の組み合わせによる字形表現も可能とするため、漢字字形を非常に少ないデータ量で記述することが可能となるだけでなく、グラフィックデザインに習熟していない一般の利用者にも漢字字形を自分で平易に作成する仕組みを実現した。この仕組みと近年注目を浴びている Web 技術である Wiki を組み合わせ、インターネット上に誰もが扱うことのできる文字データベースの構築を行った。漢字字形の多くは部品の組み合わせでできていること、および多くの異体字は元となる文字を少し変化させることで実現できるため、既存の漢字部品を操作して平易に新たな漢字字形レコードを作成できることが実証された。データベース運用開始時には外部資金（科学研究費補助金・研究公開促進費および花園大学学術フロンティア推進事業）により制作されたデータを中心とする 7,000 レコードであったものが、2 年間の運用において利用者の多数の登録により 10 万レコードを越えるものとなった。この成果は国際文字コード規格に準拠した日本最大の漢字フリーフォント「花園フォント」として公開され、複数の Linux OS でパッケージとして採用されている。

現状でデータベースに登録されている漢字字形の多くは国際標準文字コードに収録される 7 万漢字の一部であり、実際に漢字情報を扱う研究者や実務者・機関が個々に保持する外字集合を網羅するには至っていない。またデータベースのフロントエンドの英語化がほぼ完了しているにもかかわらず、利用者の多くが日本人にとどまっている。その大きな理由はデータベースの認知度が低いことが考えられる。そこで登録されている字形を実際に Web ドキュメントで活用できることを実証し、さらにデータベースの操作性を平易なものに改良するほか、日本国内外の多くの漢字字形を収集することで実用的なデータベースであることをアピールし利用の増加を促すことで、新たな漢字字形が登録されさらに実用的なデータベースになるという相乗効果を考えた。

３．研究の方法

（１）日本国内外の漢字字形資料の効率的な収集・登録

ISO/IEC 10646 規格収録漢字を中心に、東アジア圏の漢字資料に見られる漢字の収集と字形データ登録を行う。あらゆる漢字字形データを蓄積・提供することは、東洋学研究者にとどまらず、日本語・漢字学習におけるデジタル教材作成や、人名・地名表記における異体字処理において恩恵を与えるもので

あり、ひいては日本学・漢字文化の発展に寄与するものである。

（２）登録データの活用環境の実証

本研究で扱う漢字字形データベースが実用的であることを実証する目的で文字コード外ゲタ文字（=）を本来の漢字字形で表示するための機能拡張と検証を行う。具体的にはWebページの記述方法であるHTML・CSSの標準化において新たに採用されたネットワークフォントを利用し、漢字字形データベースに登録されたコード外字の集合をフォントとして目録・テキストデータベースから呼び出せるように拡張する。

（３）登録データを利用したデザイン支援機構の構築およびデータの整備

漢字字形データベースにはすでに10万レコード以上の漢字字形データが収録されている。このうち5万字弱についてはISO/IEC 10646規格の収録字である。この収録字データについて、漢字データベースプロジェクトが公開している漢字字形構造データとつき合わせることにより、どの漢字部品がどの位置にどの大きさで配置されているかを機械的に計算可能である。この情報を元にそれぞれの部品の位置・大きさの特性を数値化することで、その部品を利用して新たに漢字字形を登録する際に、推定値によってデザインしたものを提示しユーザーが微修正する形に拡張する。そのためにこの種のデータをデータベースから容易に取得できるように機能を拡張する。また、異体字が代表字と比較してどの部分に差異があるかといった異化データの整備を行い、将来的にコンピュータで異体字を容易に利用できる環境が整った以降、利用者が混乱せずに異体字を選択できる機構の基盤となるデータの整備を行う。

（４）データベースの運用と成果の公表・利用の促進

当データベースの成果を広く公表することにより、多くの利用を促し、また不足する文字種の追加登録が望めることとなる。多くの文字種が登録されればさらに有用なデータベースに成長する。この良い循環を促すために、データベースを安定的に運用し、また外部への積極的な広報を行う。

４．研究成果

（１）日本国内外の漢字字形資料の効率的な収集・登録

研究費による作業および、ボランティアの人員による作業により次のような大量の漢字字形データの収集とデータ登録を行った。①ISO/IEC 10646規格、CJK統合漢字拡張B集合（42,711字）②Unicode標準、IVD集合（2010-11-14登録分）（18,842字）この①②により、ISO/IEC 10646規格およびUnicode標準で規定される全漢字（75,619漢字および18,842異体字）を完全に収録することとなった。③和製漢字の辞典（国字収集データベース）（字形不明の2字を除く2,749字）④『国字の辞典』（飛田良文、菅原義三、東京堂出版、平成2年）（1,556字）⑤住民基本台帳ネットワーク統一文字、漢字部分（19,432字）⑥法務省戸籍統一文字（現収録54,333字）⑦住基（住民基本台帳ネットワーク）コード（現収録21,042字）⑧登記文字（登記簿使用文字）（現収録3,314字）⑨『中華字海』収録漢字のうち異体字（現収録11,935字）⑩「台湾教育部異体字字典」（現収録2,384字）⑪『古壮字字典』（現収録1,600字）⑫『東皋琴譜』（現収録884字）。以上を含め現収録字形数は313,260字となった。研究開始時点（2010年4月1日）からの増加は19万5千字強となった。

（２）登録データの活用環境の実証

データベースに登録した漢字字形をWebドキュメントから動的に呼び出し、表示するための機能を構築した。具体的にはCSSの標準化において新たに採用されたWebフォント機構を利用する。Web公開者は任意の符号位置に任意の漢字字形を割り当てたドキュメントを作成・公開できる。また、すでに公開しているドキュメントに含まれるゲタ（=）情報を、対応付けされた漢字字形に置き換える外字フィルター機能を構築した。これらの機能を活用することで、様々な漢字・異体字を含むWebドキュメントを容易に公開できる。実際に登録されている漢字グリフをWebフォントを活用することで外部の学術データベースの外字処理として利用するための実証実験が完了した。

（３）登録データを利用したデザイン支援機構の構築およびデータの整備

データベースに登録されたデータのデザイン情報を分析し、新たな漢字字形を計算によってデザイン支援する機構の研究について、そのインターフェースとなる漢字字形データの取得機構の整備と、内部データを実際の漢字字形に展開するためのプログラムの整備を行った。これらはすべてオープンデータ・オープンソースとして公開した。また漢字字形データベースに登録されているデータを活用して機械的に漢字字形を生成するプログラムのプロトタイプが完成した。このほかに、異体字関係にある2つの漢字字形のどこに差異があるかを記述するためのメタ情報を「漢字異体化データベース」としてまとめた。

（４）データベースの運用と成果の公表・利用の促進

データベースについて運用開始から５年を経て安定運用に至ることができた。また国内外の学会やメディア（特にインターネット上のコミュニティ）において当データベースについて発表・公表することで多くの利用者を導くことができた。データベースのユーザーインターフェースについては韓国語のテキストデータの準備が完了した。登録されている漢字字形データを利用して ISO/IEC 10646 規格および Unicode 標準で規定される全漢字・異体字を収録する全世界で唯一となるフリー漢字フォント「花園フォント（花園明朝)」を公開した。

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 上地宏一、明朝体漢字の異体化データベースについて、東洋学へのコンピュータ利用第２３回研究セミナー、査読無、巻名無、2012、3－31
② 上地宏一、漢字処理の現状―文字コード、フォント、外字―、三色旗、査読無、第761号、2011、3－8

〔学会発表〕（計３件）

① Taichi Kawabata, Koichi Kamichi, GlyphWiki and OpenType: A Collaborative Glyph Development Environment and its Font Exporting System, ATypI Hong Kong 2012, 2012年10月11日,InnoCentre (Hong Kong)
② 上地宏一、Web コラボレーションサービスを利用した大規模漢字集合フォントの制作、情報処理学会人文科学とコンピュータ研究会、2011 年 1 月 22 日、総合地球環境学研究所
③ 上地宏一、ウェブフォントを利用したグリフウィキ（GlyphWiki）の応用、漢字文献情報処理研究会第 13 回大会、2010 年 12 月 18 日、慶應義塾大学

〔図書〕（計１件）

① 漢字文献情報処理研究会編、好文出版、電脳中国学入門、2012、12－19、240

〔その他〕
ホームページ等
グリフウィキ　http://glyphwiki.org/
グリフウィキを活用したウェブフォント
http://fonts.jp/webfont/
花園フォント　http://fonts.jp/hanazono/

６．研究組織
(1)研究代表者
上地　宏一（KAMICHI KOICHI）
大東文化大学・外国語学部・講師
研究者番号：２０４６８７２１