2004－CH－64　(7)
2004／11／5

漢字ユビキタスを支える文字情報集積体の開発

横山詔一†[1]　笹原宏之†[1]　黒田信二郎†[2]　澤田照一郎†[3]　野島伸一†[4]　石岡俊明†[5]

†[1]国立国語研究所　〒115-8620 北区西が丘 3-9-14
†[2]紀伊國屋書店 出版部　〒150-8513 渋谷区東 3-13-11
†[3]情報処理学会 情報規格調査会 文字情報データベース開発室　〒105-0011 港区芝公園 3-5-8 機械振興会館
†[4]富士通 ミドルウェア事業本部 ミドルウェアソリューション事業部 第五開発部　〒206-8503 稲城市大丸 1405
†[5]リョービイマジクス フォントシステム部　〒114-0003 北区豊島 5-2-8

あらまし　いつでも，どこでも，だれでも，社会的に必要な漢字を使える漢字ユビキタス環境を実現するため，行政文字情報交換の基準となる「文字情報集積体（文字情報データベース）」の開発を進めている。社会的に必要な漢字の範囲を決めるには，科学的根拠のほかに国民的合意の形成が欠かせない。国立国語研究所，情報処理学会，日本規格協会の 3 者連合体は，経済産業省から委託を受けて，総務省住民基本台帳統一文字と法務省戸籍統一文字の間に存在する微妙な字形差の統一や，文字属性情報の調査に基づく情報付与作業を行い，併せて大漢和辞典の一部などの電子化を通した研究を行ってきた。また，漢字ユビキタス環境の実現を狙うという方向性そのものが国民のニーズに合致しているのかを検討するため，国立国語研究所は独自に「漢字環境学」の視点を導入して，世論調査データの解析を行った。

キーワード　漢字ユビキタス，文字情報集積体，文字グリフ，文字情報の標準化，漢字環境学

Development of national kanji database for ubiquitous computing

Shoichi Yokoyama†[1]　Hiroyuki Sasahara†[1]
Shinjiro Kuroda†[2]　Shoichiro Sawada†[3]　Shinichi Nojima†[4]　Toshiaki Ishioka†[5]

†[1]National Institute for Japanese Language
†[2]Publishing Department, Kinokuniya Co., Ltd.
†[3]Information Technology Standards Commission of Japan, Information Processing Society of Japan
†[4]Middleware Solution Division, Software Group, Fujitsu Ltd.
†[5]Font System Department, Ryobi Imagix Co.

Abstract　A comprehensive and Standardized database of kanji characters is critical to ubiquitous computing in Japan. A project team of three institute and associations, the National Institute for Japanese Language, Information Processing Society of Japan, and the Japanese Standards Association, is developing a character information database under contract with the Ministry of Economy, Trade and Industry. The database unifies the variations in the forms of kanji characters as well as includes linguistic information of the characters and the data from *Daikanwa* (大漢和辞典) kanji dictionary. Furthermore, the National Institute for Japanese Language has conducted a study on language policies regarding kanji characters in a framework called "kanji environment studies" in order to identify the opinions and the demands of language users in Japan.

Key words　ubiquitous computing, character information database, character glyph, standardization of character information, kanji environment studies

1. はじめに

必要な漢字を，いつでも，どこでも，だれでも使える･･･このような環境を，ここでは「漢字ユビキタス」という。漢字ユビキタス実現の第一歩は，文字情報交換のための標準を作成することにある。行政機関における申請・届出等の手続処理業務には，住民の姓名・住所，企業等の名称・所在地など，漢字の形を正確に確認することを求められる場合が少なくない。しかし，現状では，漢字の形を確認するための国家標準が存在しないため，申請・届出の審査に支障が生じるほか，別の行政機関と漢字データ（外字）を正確に情報交換できない。このような状況では，e-Japan 戦略や u-Japan 戦略が目指す IT 国家の実現も危うい。

国立国語研究所，情報処理学会，日本規格協会の 3 者連合体は，経済産業省からの委託を受けて，総務省住民基本台帳統一文字，法務省戸籍統一文字の電子化にかかわる文字（延べ約 8 万字）について，微妙な字形の違いなどを統一し，行政情報処理の標準となる文字集積体を構築する研究に取り組んできた。これは，図書館，公文書館，歴史資料館，郷土資料館などのデジタル・アーカイブ構築にも貢献すると期待されている。

2. 文字情報集積体の概要

2-1. コンテンツの作成

学術的検討による文字同定

国立国語研究所は，総務省や法務省から公的に提供された行政漢字データに対して，学術的な検討を施し，字体，読み，国語施策，文字コード番号など諸情報を付与した。国語施策の情報とは，「常用漢字表」の字体と，「表外漢字字体表」で示された印刷標準字体及び簡易慣用字体を指す。作業の手順は以下の通り。

(1) 総務省や法務省から提供された文字延べ約 8 万字に対して，既存の平成明朝体デザインの文字グリフ約 3 万 2,000 字種との照合作業を行った。ここでの文字グリフとは，字体の骨組みを示すための文字図形デジタルデータを指し，1 文字 1 ファイルの画像形式で Web ブラウザ等に表示させるのに利用する。

(2) 辞書に掲出されている情報を加えた。辞書に見当たらない文字については，現地の行政機関に出向いて調査を行った。

(3) 文部科学省，法務省，経済産業省などが示す諸規則を正確に適用した各種の情報を付与した。

文字グリフの補正と制作

電子政府の申請業務や省庁間の情報交換を正確に行うには，国内における行政情報で利用されている文字パターンを包括的に収集し，デザインや字形を統一する必要がある。そのために，日本規格協会は，国立国語研究所が同定・検証した文字を，スケーラブルのアウトラインデータとして作成し，文字情報集積体の字形表示用標準パターンとした。この文字グリフは平成明朝体を基にし，統一的にデザインされたきわめて高品質な文字パターン集合である。一部の文字グリフは，すでに作成済みの平成明朝体を活用または補正することで対応した。

2-2. システムの開発と運用

情報処理学会は，国立国語研究所が文字同定用に使用した「文字情報収集システム」と，一般ユーザが使用する「文字情報公開システム」の開発を担当した（これらのシステム全体を「文字情報集積体」という）。

文字情報収集システムについて

国立国語研究所で行う文字情報整理・体系化は膨大な作業を伴う。作業を円滑に進めるため，文字情報収集システムが開発された『大漢和辞典』の見出し字や読み情報のすべてを，著作権者である大修館書店と共同で電子化し，世界で初めてコンピュータに搭載した。その画面例を図 2-2-1 に示す。

文字図形統一番号→
（文字鏡番号）

住基統一文字→

戸籍統一文字→
「真」の上部「十」に注目

デザイン統一文字→
「ハ」の形状などが住基文字と微妙に異なる

大漢和辞典情報→

国語施策情報→

JIS 規格情報→

図 2-2-1　文字情報収集システムの画面例

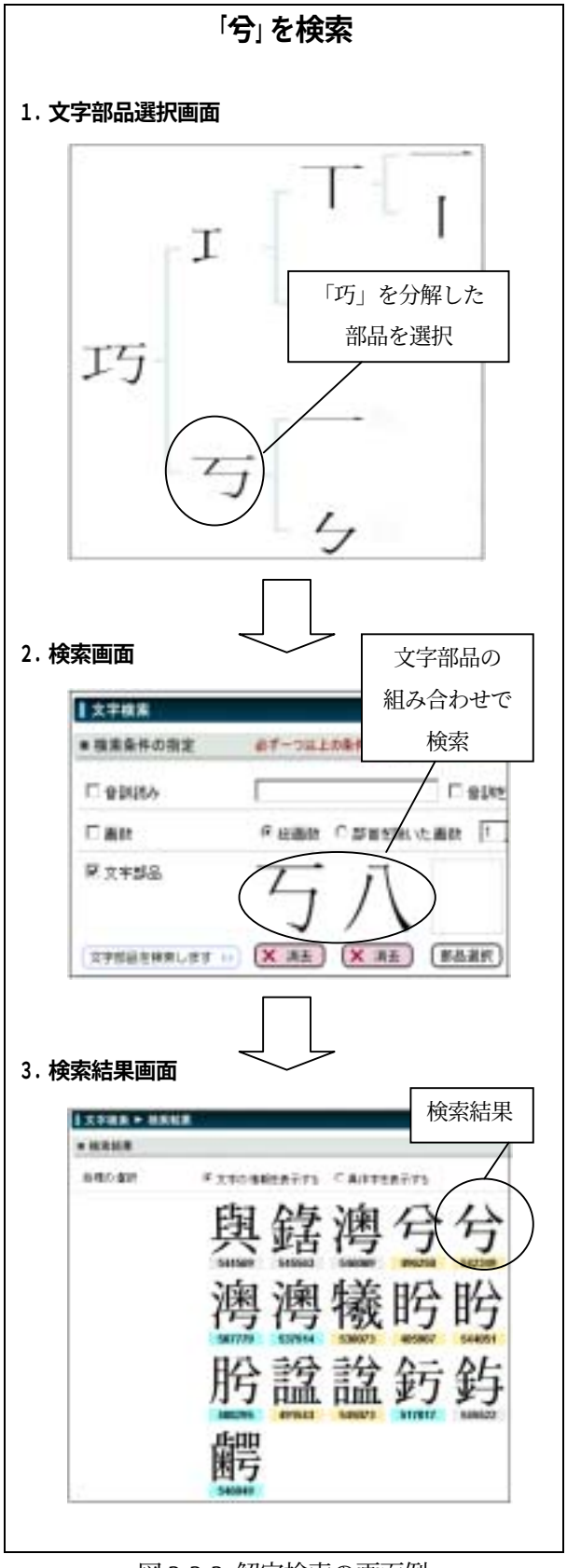

図 2-2-2 解字検索の画面例

図 2-2-3 「辺」の異体字一覧（想定例）

文字情報公開システムについて

○検索の簡便性について

各地方自治体職員や一般市民等が電子申請などにおいて利用することを視野に入れて，Web ブラウザ等で必要な文字情報を検索できるようにした。漢字の専門知識を持たない人であっても，簡便迅速に目的の文字を検索できるよう，以下の仕組みを装備している。

（解字検索機能）

部首・読みなどの特定が困難な文字については，よく知られた文字を入力し，その文字を分解して取り出した構成部品を検索キーに用いることができる。この機能を実現するため，すべての登録文字について文字の構成部品が用意されている。解字検索の例を図 2-2-2 に示す。

（関連字表示機能）

異体字関係にある文字の一覧のほか，規格内字

と異体字との異同判別に役立つ異体字マップを表示できる。その例を図 2-2-3 に示す。

○高品質な文字グリフの Web 表示について

検索画面に表示される文字は，一般市民に「なじみ」の深い明朝体（平成明朝体）でデザインした文字グリフで表示することとした。

3. 漢字環境学の導入

国立国語研究所は，現代日本で使用されている異体字について最新の調査研究を行い，それを「漢字環境学」の枠組みで整理したうえで，行政文字の字形統一や文字情報集積体の構築に応用し，文字ユビキタス社会の構築に役立てるべきだと考えている。その基本方針は，国立国語研究所プロジェクト選書№2『現代日本の異体字』（笹原・横山・ロング，2004）に明記されている。

しかし，この問題を行政情報処理の実務面だけに焦点化して議論を進めるのは，いささか視野が狭い。国民各層が漢字に対してどの程度の必要性を感じているのかをまずは明らかにした上で，漢字ユビキタス環境を構築するという目標設定が政策的に妥当なのか，国民各層のニーズに合致しているのか，などの点を確認しながら事業を進める必要があるだろう。そもそも，国民の大多数が「（将来）漢字は不要になる」という意見を持っているのであれば，国費を投入して漢字ユビキタス環境の基盤を整備する意義があまりないことになる。その場合，本プロジェクトは政策的な合理性を欠くと言わざるを得ないし，そこに国立国語研究所が参加する必要性も薄れる。諸外国の情勢を見ると，例えば韓国国立国語研究院は漢字廃止（ハングル専用）を支持する方向に動いている。

日本国民の意向を探る基礎資料を得るため，国立国語研究所は独自に「漢字環境学」の視点を導入して，世論調査データの解析を行った。

3-1. 漢字環境学とは

漢字環境学は，以下の 4 つの領域から成る。

(1) 言語領域：異体字についての国語学，日本語学，文字論，漢字学などの研究。

(2) 社会領域：新聞などマスメディアに登場した異体字を計量的に分析し，社会における異体字の使用頻度を明らかにする研究。計量国語学，社会言語学，言語政策論などの研究。

(3) 認知領域：異体字の「なじみ」や「好み」など「漢字心理」に関する研究。認知科学，認知心理学，言語心理学，心理言語学，日本語教育学，脳科学などの研究。

(4) 工学領域：異体字のユビキタス化を実現する IT 開発。情報科学，コンピュータ科学，デバイス開発などの研究。

日々の文字生活の中で，人間は自然にある漢字に接触し，その接触頻度の高低によって，その漢字に対する接触意識が生じ，それがなじみ，ひいては好みを形成すると考えられる。このような観点によるモデルが図 3-1-1 である。なお，この図には示していないが，接触頻度の要因以外に，未知の字を既知の字体との類似性判断によって渡りをつける一種の推論作用のほか，嘘字を嫌ったりする規範意識や，書体差に注意を向ける傾向が何かしら生まれたりすることによっても，字体に対する好み・なじみが影響される可能性がある。

漢字の好み・なじみは，漢字心理の一部である。漢字心理は，人間が漢字を読む（識別や包摂も含む）場合だけではなく，漢字を使用する際にも大きく影響し，それが IT 機器の利用によって社会に発信され，社会での使用頻度を変化させていく。漢字環境学は，結局のところ，図 3-1-1 の全体をカバーする学問である。以下，認知領域，社会領域，工学領域の研究例を紹介する。

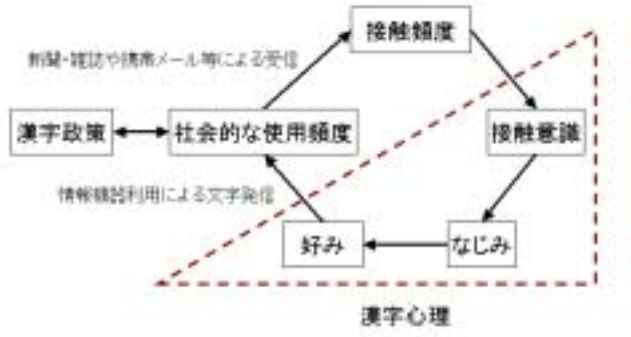

図 3-1-1　漢字環境の諸要素

3-2. 認知領域：漢字心理の調査研究例

漢字に関する２つの世論調査の結果から，文字生活における国民の漢字心理を推測した。図 3-2-1 は，文化庁国語課が行った世論調査の結果とインターネットを活用した Web 調査の結果を比較したものである。

国民一般の漢字心理

文化庁国語課は，国語施策の参考にするため，毎年全国規模で「国語に関する世論調査」を実施している。図 3-2-1 は 2002 年の 11 月 14 日から 12 月 2 日にかけて行われた漢字に関する意識調査の結果である。調査対象は全国の 16 歳以上の男女 3,000 名で，個別面接調査法によってデータを収集し，有効回収数（率）は 2,200 名（73.3%）であった（文化庁国語課，2003）。表中の質問項目は，パーセントの高いものから順に並べた（8 個の選択肢の中からの複数選択）。

文化庁世論調査のデータからは,「漢字を覚えるのは大変なので，なるべく使わない方がよい」や「ワープロなどがあるので，これからは漢字を書く必要は少なくなる」と考えている人は全国民の 3〜4%程度であって，最下位に位置付くことが分かった。対照的に,「日本語の表記に欠くことのできない大切な文字である」は第 1 位で，70%以上の支持を集めている。

デジタル先進派の漢字心理

では，インターネットを使いこなしている人の漢字心理はどうなのだろうか。今後の情報社会をリードするであろう，このようなデジタル先進派は，漢字についてどのような意見を持っているのだろうか。漢字は「古くさい」というイメージを抱き，心理的に敬遠しているのだろうか。

2004 年 2 月下旬に, インターネットを活用した Web 調査により，20 歳以上の女性約 500 名を対象にデータを収集してみた。調査サンプル（標本）は，日本全国 12 万人のパネル（調査協力者）から無作為（ランダム）に抽出したものである。（回収結果は, 20 歳代, 30 歳代, 40 歳代それぞれ 120 名ずつ，50 歳代は 102 名，60 歳代が 50 名。合計 512 名。地域は，新潟県，東京都，埼玉県，千葉県，神奈川県，愛知県，大阪府，京都府，兵庫県。データ収集は，Web 調査の実績が豊富なインフォプラント社に委託。）なお，Web 調査サンプルの生活様式（ライフ・スタイル）は，一般の人よりも「やや先進的」であることが，事前の社会学的な調査などにより，あらかじめ明らかになっている（横山，2004）。

Web 調査の被調査者は，電子機器による漢字変換の利便性を十分に享受しているので，漢字学習の必要性を一般人（文化庁世論調査）よりも低く感じているのだろうか。これらの点を確認するため，先の文化庁による世論調査とまったく同じ質問項目を呈示した。

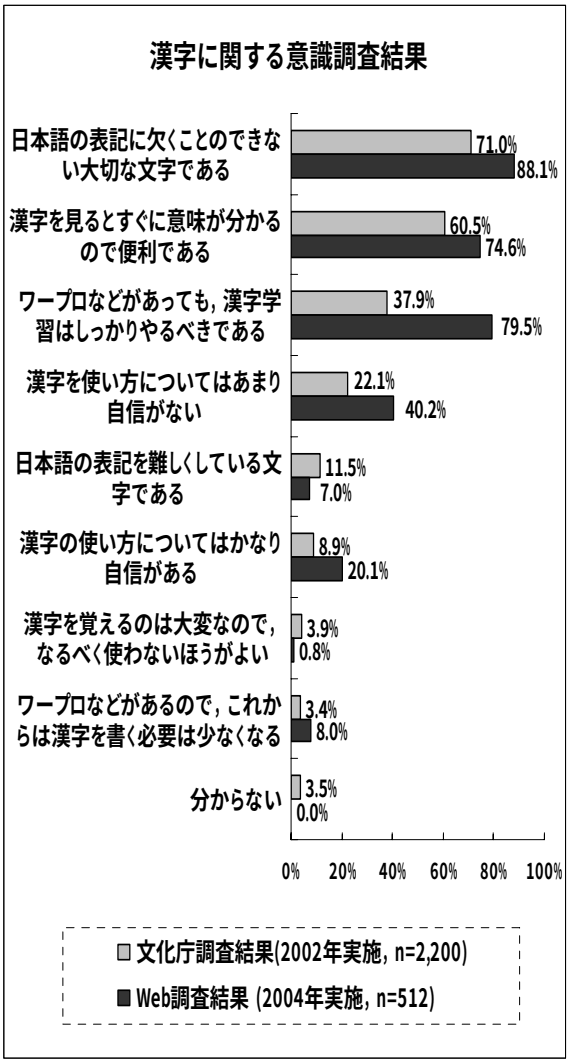

図 3-2-1 漢字心理の調査結果

Web 調査の結果を図 3-2-1 に示す。全体的な傾向として,「先進的な人は，一般の人よりも漢字の重要性を強く認識している」と言えるだろう。「漢字を覚えるのは大変なので，なるべく使わない方がよい」を選んだ人は 1%以下であった。逆に,「ワープロなどがあっても，漢字学習はしっかりとやるべきである」は文化庁世論調査の数値よりも 40%以上も高くなっている。さらに,「日本語の表記に欠くことのできない大切な文字である」は 90%近くを占めており，漢字に否定的な意見を圧倒していた。

3-3. 社会領域：官報外字の調査研究例

国立国語研究所は国立印刷局と『官報』を対象にした共同研究を行っている。これまでに,『官報』で使用された外字のうち延べ 16,289 字の調査を行った。これらは，過去に使用された外字の全部ではなく，最近の一定期間内に国立印刷局が試験的に収集したデータである（以下，これを官報外字という）。

官報外字のうち，同じ文字が別の外字番号を与えられている「重複文字」を調査したところ，延べで約 800 字あった。これは調査対象としたデータの全体の約 5%に達し，目視による重複チェックが困難であったことを示唆している。重複文字を除く異なり字数は約 15,500 だった。

次に，住民基本台帳統一文字と戸籍統一文字を合体させた文字図形統一番号（図 2-2-1 を参照）と，官報外字との対応表を作成した。その結果，文字図形統一番号のカバー率（ヒット率）は約 20%であった。これは，現代日本の異体字が予想以上に複雑な様相を呈している証左といえよう。文字図形統一番号でカバーできなかった文字の例を図 3-3-1 に示す。なお, 2004 年 8 月における『今昔文字鏡』のカバー率は 65.8%であった。

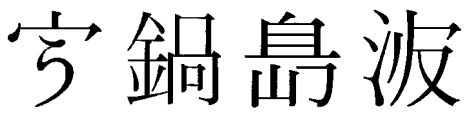

図 3-3-1 文字図形統一番号を持たない字の例

3-4. 工学領域：日本語 Web 入力方式の開発例

漢字ユビキタス環境の実現は，日本語ユビキタス環境の構築にもつながる。国際社会における日本語の地位を高めるため，国立国語研究所は，世界のどこでも日本語を Web で入力できるシステムの研究を進めている。

JiBOOKS と JiWORDS

早稲田大学図書館などの蔵書情報を，海外のブラウザから日本語で検索できるようにするため，国立国語研究所は「JiBOOKS」（注 1）を開発した。（注 1）http://btonic.est.co.jp/jibooks/78jis/

このシステムは，日本語を Web で入力できる IME を搭載している。この WebIME は，日本語環境のないブラウザでも，インターネットを介して利用できる。日本語をローマ字で入力し（図 3-4-1），変換ボタンをクリックすると漢字単語などの文字列に変換できる（図 3-4-2）。

国際交流基金と共同でマレーシアの日本語教師や日本語学習者約 250 名を対象に実施したモニター調査の結果によると，JiBOOKS のようなシステムに対して高い必要性があることが示されている（Yokoyama, Lee, & Ishida, 2004）。この状況などを踏まえて，国立国会図書館は，2004 年中に英語版 OPAC トップページから JiBOOKS に直接リンクをはる。また，国語辞典などを海外に提供するシステム「JiWORDS」（注 2）にも，この WebIME が搭載され，試験的に運用されている。（注 2）http://btonic.est.co.jp/JiDic/

図 3-4-1 JiBOOKS ローマ字入力画面

図 3-4-2 漢字変換の例

Interstage Charset Manager（注 3）

富士通は行政情報処理における外字問題の解決に取り組み, 日本語 Web 入力システムの商品化に成功している。(注 3)http://interstage.fujitsu.com/jp/output/charsetmgr/index.html

従来の技術では，外字を入力・表示する際は，国内であっても，クライアントに日本語資源（外字フォントや辞書など）を配布する必要があった。これは煩雑な作業である。この問題を解消するため,「Interstage Charset Manager Web 入力」は，サーバに日本語資源をすべて管理させ，クライアントはブラウザのみを搭載する方式を採用した。これにより，パソコンにおけるローマ字仮名漢字変換と同じ一連の動作，つまり，ローマ字や仮名で読みを入力し，変換キーで漢字を選択するやり方で外字を簡便に入力・表示することが可能になり，操作性が格段に向上した（図 3-4-3）。

図 3-4-3 Interstage Charset Manager の例

4. まとめ

世論調査の結果から，漢字ユビキタス環境の構築は国民各層の支持を得られると予測できる。(韓国とは事情が異なるように見える。)

社会的に必要な漢字の範囲を決めるには，科学的根拠のほかに国民的合意の形成が欠かせない。その点で，国立国語研究所＋情報処理学会＋日本規格協会の 3 者連合体制が確立したという事実は重要な意味を持つと言えよう。

引用文献（アルファベット順）

文化庁文化部国語課（2003）『平成 14 年度　国語に関する世論調査〔平成 14 年 11 月調査〕』，文化庁

日本規格協会・国立国語研究所・情報処理学会（2004）『平成 15 年度 経済産業省委託 汎用電子情報交換環境整備プログラム成果報告書』，日本規格協会

笹原宏之・横山詔一・エリク＝ロング〔著〕（2003）『現代日本の異体字――漢字環境学序説――』国立国語研究所プロジェクト選書№2，三省堂

横山詔一（2004）「文字処理の認知科学」月刊『言語』8 月号「特集 言語にとって文字とは何か」pp.56-63，大修館書店

Yokoyama S., Lee S. L., & Ishida, T. (2004) Bibliographic catalogue web-based search system designed for non-Japanese browsers "JiBOOKS" : Report on evaluation survey in Malaysia, The National Institute for Japanese Language

Yokoyama S., Long E., Yoneda J., Wada Y., Kuroda S., & Shimokawa K. (2004) Web IME: Web-based Japanese input method editor applied to a search system for library catalogues, IPSJ SIG Technical Report, 2004-DD-46 (7), pp.43-47

附記

本研究の前半部は，経済産業省委託研究「汎用電子情報交換環境整備プログラム」の成果の一部である。後半部は，文部科学省科学研究費補助金（基盤(C)(2)，課題番号 16520290，研究代表者：横山詔一）などによる。

このプロジェクトは, 大阪府立大学名誉教授・樺島忠夫先生, 慶應義塾大学教授・石崎 俊先生，経済産業省・堀坂和秀様, 文化庁国語課・氏原基余司様，日本規格協会・若井博雄様，堤伸介様, 中野誠司様, 情報処理学会・三田真弓様, 日立製作所・荒木幸治様，ネクストソリューション・長村 玄様，文字鏡ネット・谷田貝常夫様，文字鏡研究会・古家時雄様，紀伊國屋書店・有馬由紀子様，国立国語研究所・エリク＝ロング様，米田純子様，和田志子様，澁谷朋子様ほか多くの方々のご支援によって進められた。記して感謝の意を表する。